THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Computing abundance constraints in *Saccharomyces cerevisiae*'s metabolism

BENJAMÍN JOSÉ SÁNCHEZ

Department of Biology and Biological Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2019

Computing abundance constraints in *Saccharomyces cerevisiae*'s metabolism

BENJAMÍN JOSÉ SÁNCHEZ

ISBN 978-91-7597-863-5

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 4544
ISSN 0346-718X

Division of Systems and Synthetic Biology
Department of Biology and Biological Engineering
Chalmers University of Technology
SE-412 96 GOTHENBURG
Sweden
Telephone: + 46 (0) 31 772 10 00

Cover: Abundance constraints in yeast metabolism.

Printed by Chalmers Reproservice,
Gothenburg, Sweden 2019

Computing abundance constraints in *Saccharomyces cerevisiae*'s metabolism

Benjamín J. Sánchez
Department of Biology and Biological Engineering
Chalmers University of Technology

Abstract

The unicellular eukaryotic organism *Saccharomyces cerevisiae* (budding yeast) is routinely used for production of high-value chemical compounds in the biotechnology industry. To improve production yields, it is fundamental to understand cellular metabolism, i.e. all biochemical reactions that occur inside the cell. In the past 20 years, genome-scale metabolic models (GEMs) have risen as computational tools for simulating all possible metabolic phenotypes that the cell can attain, while respecting constraints such as mass balances and reaction reversibilities. However, the number of metabolic states bound to only those constraints is infinite; therefore, it becomes necessary to include additional condition-specific constraints. Moreover, we would like these constraints to reflect physical limitations inside the cell, avoiding arbitrary *ad-hoc* bounds.

In this thesis, approaches for including abundance constraints (i.e. constraints based on absolute abundances of different biomolecules) are evaluated in a GEM of *S. cerevisiae*. First, the GEM approach and how it has been used in *S. cerevisiae* is reviewed, identifying key areas for development. Afterwards, the concepts of sustainable model development and multi-layer experimental data generation are presented as foundation stones for constructing integrative analysis. Regarding the first concept, a systematic way of recording changes in a GEM using a version-controlled system is introduced, allowing reproducibility and open collaboration from the community. Regarding the second concept, a multi-omics dataset of yeast grown under different temperature, osmotic and ethanol stresses is presented and used throughout the thesis for studying metabolism.

The major part of this work focuses on the integration into GEMs of abundance data of two types of bio-molecules: lipids and enzymes. First, a method for integrating lipid requirements in an unbiased way (SLIMEr) is presented and implemented for yeast, to show that lipid metabolism can be re-arranged without spending high amounts of energy. Secondly, a method for adding so-called "enzyme constraints" into a GEM (GECKO) is developed. These enzyme constraints limit reaction rates by the absolute abundance of enzymes, and prove to be crucial for explaining yeast physiology and computing enzyme usage in metabolism. Thirdly, the quantification technique used for estimating enzyme abundances is analyzed in terms of accuracy and precision, and further improved by varying the normalization and scaling steps. Finally, GECKO is used on the stress dataset to create enzyme-constrained models of yeast representing each stress condition. This allows comparing the distribution of enzyme usage within and between conditions, highlighting enzymes that play an important role in the metabolic response to stress.

Keywords: Genome-scale modeling, flux balance analysis, proteomics, lipidomics.

List of Publications

This thesis is based on the work contained in the following papers and manuscripts:

Paper I: *Genome scale models of yeast: Towards standardized evaluation and consistent omic integration*. Benjamín J. Sánchez, Jens Nielsen. Integrative Biology (2015). 7(8): 846–858.

Paper II: *Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast*. Petri-Jaan Lahtvee, Benjamín J. Sánchez, Agata Smialowska, Sergo Kasvandik, Ibrahim Elsemman, Francesco Gatto, Jens Nielsen. Cell Systems (2017). 4:1-10.

Paper III: *SLIMEr: probing flexibility of lipid metabolism in yeast with an improved constraint-based modeling framework*. Benjamín J. Sánchez, Feiran Li, Eduard J. Kerkhoven, Jens Nielsen. BMC Systems Biology (2019). 13(4).

Paper IV: *Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints*. Benjamín J. Sánchez, Cheng Zhang, Avlant Nilsson, Petri-Jaan Lahtvee, Eduard J. Kerkhoven, Jens Nielsen. Molecular Systems Biology (2017). 13(8): 935.

Paper V: *Benchmarking accuracy and precision of intensity-based absolute quantification of protein abundances in Saccharomyces cerevisiae*. Benjamín J. Sánchez, Petri-Jaan Lahtvee, Kate Campbell, Sergo Kasvandik, Rosemary Yu, Iván Domenzain, Aleksej Zelezniak, Jens Nielsen [manuscript].

Paper VI: *Limitations on enzyme usage in Saccharomyces cerevisiae's metabolic stress response*. Benjamín J. Sánchez, Petri-Jaan Lahtvee, Iván Domenzain, Eduard J. Kerkhoven, Jens Nielsen [manuscript].

Additional papers and manuscripts not included in this thesis:

Paper VII: *The RAVEN Toolbox 2.0: a versatile platform for metabolic network reconstruction and a case study on Streptomyces coelicolor*. Hao Wang, Simonas Marcišauskas, Benjamín J. Sánchez, Iván Domenzain, Daniel Hermansson, Rasmus Agren, Jens Nielsen, Eduard Kerkhoven. PLoS Computational Biology (2018). 14(10): e1006541.

Paper VIII: *Memote: A community-driven effort towards a standardized genome-scale metabolic model test suite*. Christian Lieven, Moritz Emanuel Beber, Brett G. Olivier, Frank T. Bergmann, Meric Ataman, Parizad Babaei, Jennifer A. Bartell, Lars M. Blank, Siddharth Chauhan, Kevin Correia, Christian Diener, Andreas Dräger, Birgitta Elisabeth Ebert, Janaka N. Edirisinghe, Jose P. Faria, Adam Feist, Georgios Fengos, Ronan M. T. Fleming, Beatriz Garcia-Jimenez, Vassily Hatzimanikatis, Wout van Helvoirt, Christopher Henry, Henning Hermjakob, Markus J. Herrgard, Hyun Uk Kim, Zachary King, Jasper Jan Koehorst, Steffen Klamt, Edda Klipp, Meiyappan Lakshmanan, Nicolas Le Novere, Dong-Yup Lee, Sang Yup Lee, Sunjae Lee, Nathan E. Lewis, Hongwu Ma, Daniel Machado, Radhakrishnan Mahadevan, Paulo Maia, Adil Mardinoglu, Greg L. Medlock, Jonathan Monk, Jens Nielsen, Lars K. Nielsen, Juan Nogales, Intawat Nookaew, Osbaldo Resendis, Bernhard Palsson, Jason A. Papin, Kiran Raosaheb Patil, Mark Poolman, Nathan D. Price, Anne Richelle, Isabel Rocha, Benjamín J. Sánchez, Peter Schaap, Rahuman S. Malik Sheriff, Saeed Shoaie, Nikolaus Sonnenschein, Bas Teusink, Paulo Vilaca, Jon Olav Vik, Judith A. Wodke, Joana C. Xavier, Qianqian Yuan, Maksim Zakhartsev, Cheng Zhang. bioRxiv 350991; doi: https://doi.org/10.1101/350991.

Paper IX: *Yeast evolved to have excess glycolytic capacity at low growth rates*. Jianye Xia, Benjamín J. Sánchez, Yu Chen, Kate Campbell, Sergo Kasvandik, Jens Nielsen [manuscript].

Paper X: *Yeast 8: The consensus genome-scale model of yeast as a standard for community development and multi-scale simulation of metabolism*. Hongzhong Lu, Feiran Li, Benjamín J. Sánchez, Zhengming Zhu, Gang Li, Iván Domenzain, Simonas Marcišauskas, Petre Mihail Anton, Dimitra Lappa, Christian Lieven, Moritz Emanuel Beber, Nikolaus Sonnenschein, Eduard J Kerkhoven, Jens Nielsen [manuscript].

Paper XI: *High-throughput metabolic reprogramming using gRNA libraries coupled with CRISPRa and malonyl-CoA biosensor*. Raphael Ferreira, Christos Skrekas, Alex Hedin, Benjamín J. Sánchez, Jens Nielsen, Florian David [manuscript].

Paper XII: *GECKO 2.0: A generalized toolbox for enhancing genome-scale metabolic models with enzymes constraints*. Iván Domenzain, Benjamín J. Sánchez, Eduard J Kerkhoven, Jens Nielsen [manuscript].

Contribution Summary

Paper I: I reviewed the literature and wrote the original manuscript.

Paper II: I performed the analysis for estimating degradation rates and contributed with computational simulations and other data analysis.

Paper III: I designed the study, co-implemented the mathematical formulation, developed the computational method, performed the computational simulations, analyzed the results and wrote the original manuscript.

Paper IV: I co-designed the study, co-implemented the mathematical formulation, co-developed the computational method, performed the computational simulations, analyzed the results and wrote the original manuscript.

Paper V: I designed the study, performed the data analysis and wrote the original manuscript.

Paper VI: I designed the study, performed the computational simulations, analyzed the results and wrote the original manuscript.

Paper VII: I contributed with software development.

Paper VIII: I contributed with implementation and validation of the software.

Paper IX: I performed the analysis on enzyme usage.

Paper X: I co-developed the software and performed some of the model analysis.

Paper XI: I performed the computational simulations for predicting gene targets.

Paper XII: I co-designed the study and co-developed the software.

Preface

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Biology and Biological Engineering at Chalmers University of Technology. The PhD studies were carried out between October 2014 and March 2019 at the division of Systems and Synthetic Biology (SysBio) under the supervision of Jens Nielsen. The project was co-supervised by Eduard Kerkhoven and examined by Stefan Hohmann. It was mainly funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 686070, the Knut and Alice Wallenberg Foundation, and the Novo Nordisk Foundation. Financial support from CONICYT (grant #6222/2014) is also acknowledged.

Benjamín J. Sánchez

March 2019

Table of Contents

Abbreviations

ATP	Adenosine triphosphate
CL	Cardiolipin
CO_2	Carbon dioxide
DAG	Diglyceride
DNA	Deoxyribonucleic acid
FAME	Fatty acid methyl ester
FBA	Flux balance analysis
FFA	Free fatty acid
FVA	Flux variability analysis
GAM	Growth associated ATP maintenance
GC	Gas chromatography
gDW	Grams of cell dry weight
GECKO	Genome-scale modeling with enzyme constraints, using kinetics and omics
GEM	Genome-scale metabolic model
GUI	Graphical user interface
HPLC	High-performance liquid chromatography
iBAQ	Intensity-based absolute quantification
LC	Liquid chromatography
MS	Mass spectrometry
NaCl	Sodium chloride
NGAM	Non-growth associated ATP maintenance
mRNA	Messenger ribonucleic acid
O_2	Oxygen
PC	Phosphatidylcholine
PCA	Principal component analysis
PE	Phosphatidylethanolamine
pFBA	Parsimonious flux balance analysis
PI	Phosphatidylinositol
pO_2	Partial pressure of oxygen
PS	Phosphatidylserine
RNA	Ribonucleic acid
RPM	Revolutions per minute
SBML	Systems biology markup language
SE	Sterol ester
SILAC	Stable isotope labeling by amino acids in cell culture
SLIME	Split lipid into measurable entities
SysBio	Division of systems and synthetic biology
TAG	Triglyceride
TCA	Tricarboxylic acid
TPA	Total protein approach

Captain, the most elementary and valuable statement in science, the beginning of wisdom, is "I do not know". I do not know what that is sir.

LCdr. Data, Star Trek TNG, S02E02

Acknowledgments

Acknowledgments are probably the most read section of any PhD thesis [*citation needed*]. For me, they have also been the most fun section to write. During my PhD studies I have had a fantastic support network at the academic, instrumental and emotional levels [1], and I wish to express my gratitude to everyone that has been a part of this journey. So here it is: Thanks.

Thanks to my supervisor, Jens, for inviting me to be a part of the SysBio family, for challenging me with hard questions, for always supporting my decisions, for patiently waiting for my delayed manuscripts, and for always finishing meetings with a smile. You are a great mentor and I always look forward to our next talk. Thanks also to my co-supervisor, Ed, for lengthy discussions, helpful suggestions and thorough reviews. I hope we continue to collaborate on great projects in the coming years.

During my studies I have had the honor to join forces with many great researchers and even better people. Thanks Petri for inviting me to collaborate from the start, and for generating almost all of the data I have analyzed in this thesis; without you this work would have not been possible. Thanks Avlant for your unbridled enthusiasm and providing me with both guidance and friendship throughout the years. Thanks Kate for all those coffee "breaks" discussing proteomics; your attention to detail is amazing. Thanks to my *frenduru* Raphaël for the science and non-science talks, looking forward to harvesting the fruits of our work. Finally, thanks Iván for being both a great office mate and co-developer, and for our never-ending talks; I'm sure someday we will figure out the shape of the flux cone.

Thanks to all of the SysBio group, where I have shared and science'd for the past four years. Thanks Cheng for our initial work on formulating GECKO. Thanks Feiran for your motivation and great discussions that lead to SLIMEr. Thanks Hongzhong, Demi, Mihail and Simonas for your help in developing *yeast-GEM*. Thanks Martin, Michi, Eugene, Gang, Yu, Ibrahim, Carl and all the attendees of the yeast subgroup meetings, for your valuable feedback and your patience with my frequent interruptions. Thanks Pouyan, Amir, Jianye, Rosemary, Hao, Stefan T, Sergo K and Aleksej for your great input in the different works included in this thesis. Thanks Olena, Gheorghe and Ievgeniia for present and future collaborations. Thanks Martina, Erica, Josefine, Anne-Lise and all the administrative staff for answering all of my questions. Thanks to all research engineers both in the wet-lab and dry-lab for making of SysBio a well-oiled machine.

I am deeply thankful of my collaborators in DTU, who guided me when developing many of the tools presented in this thesis. Thanks Henning for your work translating my ideas into useful software. Thanks Christian for opening my eyes to model development. Thanks Moritz for being my go-to guy whenever things would break. Thanks Niko for all the support

and encouragement. Additionally, thanks to all of the members of the DD-DeCaF project for inspiring discussions and great meetings in the past three years.

Life as a PhD student has entailed many challenges outside research. Thanks Bella, Alex P, Anna P, Giulio, Cecilia, David N, Promi, Gaowa and many others in the PhD student council for fighting alongside me for student rights. Thanks Marie, Shaq, Xin and all previous and current members of the core-value group for working together to achieve a better working environment for SysBio.

These four years have been extremely tough but also extremely fun: Thanks to my running buddies (Ausra and Jenny) and the ever-changing climbing group, for keeping me sane enough to complete my studies. Thanks Anastasia, Verena and George for giving me something to dream about every Wednesday. Thanks Elias for introducing me to Sweden. Thanks Ximena for being a second mother when I started. Thanks Florian for always making me laugh. Thanks Bella for your amazing energy. Thanks Yasi, Markus J and Jeroen for all the real talks. Finally, thanks to José, Clara, Mark B, Tatiana, Julia, Gatto, Leif, Alex B, Sakda, David B, JC, Parizad, Sylvain, Leonie, Anna W, Francesca, Saki, Jon, Christoph and so many other beautiful souls from SysBio that I have met during my time here. It has really been amazing to be exposed to so many cultures, and I leave grateful and knowing that wherever I will be in the world, I will know someone that I can annoy.

As an expat, I've been #blessed with a ragtag bunch of misfits that I see as my Swedish family: thanks Josh, Helen, Eric, Paulo, Cat and Marlous for being there through thick and thin. I know we won't lose contact, after all we have a business to start! Thanks also to Bill, Leslie, Niklaus, Blantan, Debby and DMM for entertaining me to tears every week and allowing Remmi to be the best version of himself. Finally, thanks again to Cat, who endured my whims and endless chatter for the past three years as my flat mate. I hope it was not too bad.

The final words here are for my family. Mom and dad, your support throughout the years every time I did not believe in myself will never stop inspiring me. Trini, Montse, Rosa, Iñigo, Clemente and Emilio, you are stars that shine even in my darkest nights and for that I am eternally grateful. Jaime and Gustavo, I'm thrilled that you've joined our family. Andrés and Aurora, your arrival to this world has filled me with joy. Every day I think of the next time I'll get to see you guys, I love you.

Benja

1. Background

Biotechnology, the exploitation of biological processes for the benefit of humanity, has become one of the main industries in our current economy [2]. A big part of biotechnology relies on studying microorganisms, and how to tune them to best suit the corresponding application. Metabolism, the interplay of chemical reactions inside microorganisms, is a key layer for understanding microorganisms' responses and for using them to our advantage. In this chapter I introduce yeast metabolism (my subject of study in this thesis), go through **Paper I**, which reviews metabolic modeling in yeast, and introduce the aim and significance of this thesis.

1.1. *Saccharomyces cerevisiae*

Saccharomyces cerevisiae, also referred to as baker's yeast or budding yeast, was most likely the first microorganism ever used for a biotechnology application: alcoholic fermentation [3]. Until today, it is one of the most utilized organisms in industrial biotechnology [4], with applications ranging from bread, beer and wine making, to production of biofuels, food additives and pharmaceuticals [5]. Note that in this thesis, the word "yeast" will be often used to refer to *S. cerevisiae*, with apologies to other yeast species [6].

S. cerevisiae has two important properties that makes it an interesting subject of research: ease of growth and high complexity. On one hand, it is a unicellular organism that grows fast in laboratory conditions, which eases the generation of data compared to e.g. plant or human cells. However, unlike other fast-growing organisms such as *Escherichia coli* and *Bacillus subtilis*, it is eukaryal, meaning among others that cellular functions are compartmentalized into several organelles (nucleus, mitochondria, etc.) just like higher organisms such as human cells. Therefore, *S. cerevisiae* can also be used as a model organism for studying complex biomolecular processes, such as human diseases [7]. In summary, budding yeast is of high value both for the industrial and research communities; therefore, a better understanding of its inner workings would yield economic benefits and fundamental insight into the biological process of life.

1.2. Cellular metabolism

Metabolism, the process of generating energy and components for growth, is arguably the most important cellular process [8]; without it the cell would not be able to generate energy to perform any other basic task required for life, such as reproduction or homeostasis. Metabolism consists of thousands of different chemical reactions, which, for simplicity, are

classified into different metabolic pathways. These reactions convert nutrients into thousands of different intermediate chemical compounds, referred to as metabolites, some which are needed for growth and reproduction, and some which are excreted as by-products (**Figure 1**). As most cellular reactions would either not occur spontaneously due to their thermodynamic properties, or would be too slow to support life in a competitive environment with limited resources, cells express enzymes, a type of protein which can catalyze reactions, i.e. accelerate their rate. Different enzymes are specific to different metabolic reactions, and the required information to express them is encoded in the cell's DNA.

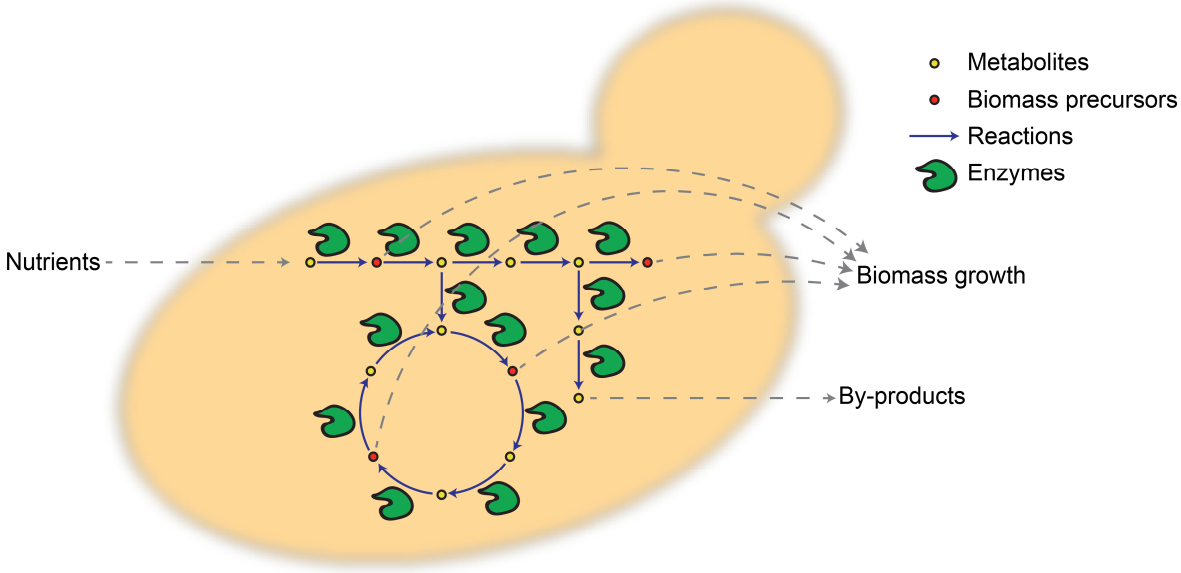

Figure 1: A simple diagram of metabolism.

Metabolism can broadly be classified into catabolism and anabolism [9]. Catabolism includes all metabolic reactions that break down metabolites to generate energy, most often in the form of adenosine triphosphate (ATP). Anabolism in turn includes reactions that build up biomolecules that form the main components of the cell: proteins, RNA, DNA, carbohydrates and lipids. The first four mentioned components are polymers, i.e. large molecules made up of a combination of a few types of small molecules (monomers). Proteins, RNA and DNA are long sequential combinations of amino acids, ribonucleotides, and deoxyribonucleotides, respectively, and carbohydrates are sequential (but sometimes branched) combinations of monosaccharides. Therefore, a central part of anabolism is forming these monomers, or "building blocks" of the cell, which will later be assembled into functional components.

Metabolism can be tuned by the cell to adjust its needs, depending on the environmental conditions. In this regard, the field of metabolic engineering [9] has emerged in the past 30 years as a way of studying these metabolic decisions, and how they can be modified to favor production of chemical compounds with economic value, which otherwise would have to be synthetized using less sustainable practices, such as chemical synthesis. Making these modifications has become rather straightforward, thanks to recent techniques of gene editing [10] that allow fine-tuning the expression of any gene inside the cell to control the level of enzyme. However, it remains challenging to understand at the systems-level how yeast uses

its metabolism for growth and production of metabolites. In this thesis, I study metabolism of yeast from a holistic perspective, i.e. considering all pathways, reactions and metabolites at the same time.

1.3. Genome-scale modeling of metabolism: Starting from the bottom

To properly understand metabolism, we need the aid of computers. This need arises from our brains' limited capacity to process information, an observation referred to as Miller's law [11]. As metabolism consists of thousands of reactions and metabolites, we must rely on a computational representation to understand it on a systems level. In this regard, the field of systems biology [12] has emerged as the formal study of complex biological systems, using mathematical modeling and high-throughput data. In this section I review genome-scale metabolic modeling, a "bottom-up" approach and the main technique for modeling metabolism in systems biology, together with the challenges that the field presents.

1.3.1. Mechanistic models of metabolism

Modeling metabolism has always been at the core of systems biology [12]. Already when the biochemical steps of the main metabolic pathways were being elucidated, mathematical models of metabolism were being developed for simulation and analysis. These models can broadly be classified in phenomenological and mechanistic models [13]. Phenomenological models, also referred to as "top-down" models, are constructed from experimental measurements of cellular information (oftentimes at a genome-wide level), and in general aim to discover previously unknown molecular interactions, using the power of statistics. Instead, mechanistic models are built from our understanding of singular components and their interactions between them, to form in a "bottom-up" approach a working model of the cell. In this thesis, I focus on mechanistic models of metabolism.

Two main formalisms have been developed for mechanistic modeling of metabolism: kinetic modeling and stoichiometric modeling [14] (**Table 1**). In kinetic models, reaction rates, also referred to as fluxes, are modeled as a function of metabolite concentrations (typically in g/L units), using previously inferred mathematical representations of reaction mechanisms, and metabolite concentrations are modeled as a function of time, using ordinary differential equations. In stoichiometric models, sometimes referred to as constraint-based models [15], a pseudo-steady state is instead assumed, which dictates that under short timescales the accumulation of intracellular metabolites can be neglected [9]. Metabolite concentrations are therefore not modeled, and reaction fluxes (typically in mmol/gDWh units) are inferred by imposing steady-state mass balances on each metabolite. In this thesis, I explore the stoichiometric modeling approach, and its use for elucidating yeast physiology.

Table 1: Differences between kinetic and stoichiometric modeling. M is the set of metabolite concentrations, k the set of parameters needed *a priory*, v the metabolic fluxes, and LB and UB the lower and upper bounds, respectively, of the metabolic fluxes.

	Kinetic Modeling	Stoichiometric modeling
Type of model	Dynamic	Stationary
Unknown variable	Concentrations: M [g/L]	Fluxes: v [mmol/gDWh]
Mass balances	$\frac{dM}{dt} = S \cdot v(M, k)$	$S \cdot v = 0$
Additional constraints	$M(t = 0) = M_0$	$LB \leq v \leq UB$
Number of parameters	Several	Few
Number of solutions	Single	Infinite
Linearity	Non-linear	Linear
Typical size of network	<100 metabolites	>1000 metabolites
Computational time required	~minutes	~milliseconds

A main advantage of stoichiometric models is their computational efficiency, which is due to their linear structure. By assuming steady state, mass balances for each metabolite yield simple linear relationships between metabolic fluxes: the sum of all fluxes that produce a given metabolite must be equal to the sum of all fluxes that consume it. This can be expressed as a simple equation in which fluxes, represented by a vector, are multiplied by a matrix known as the stoichiometric matrix (**Figure 2**). The columns of this matrix indicate the stoichiometry of reactions, and the rows indicate the mass balances for each metabolite. The only remaining requirement for simulations is to impose inequality constraints on fluxes, based on either measured data (e.g. the uptake of nutrients from the media) or known impossibilities (e.g. irreversible reactions) (**Figure 2**). In turn, kinetic models are non-linear, and typically take much longer to simulate [16] (**Table 1**).

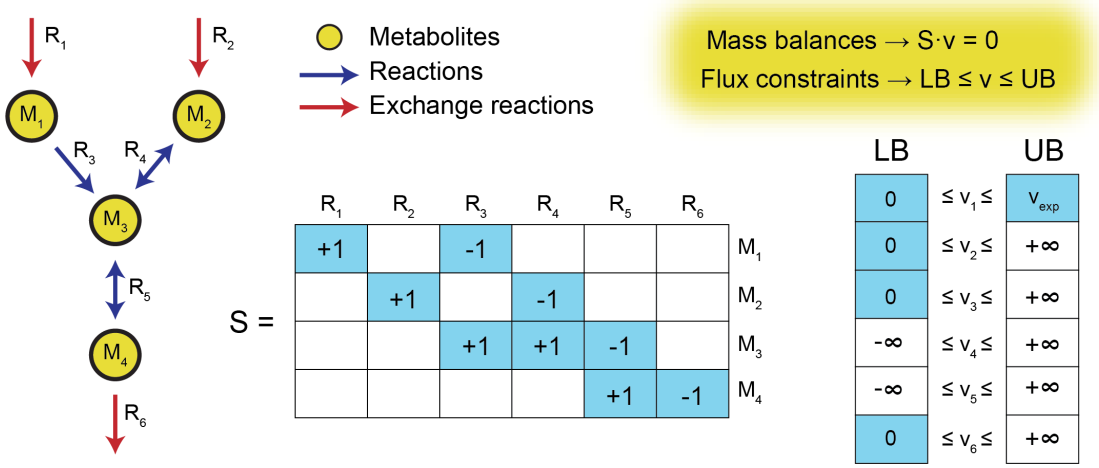

Figure 2: Example of stoichiometric modeling for a toy network. M_i are metabolites, R_i are reactions, v_i are reaction fluxes, S is the stoichiometric matrix, and LB and UB are lower and upper bounds, respectively. Note that reactions R_4 and R_5 are reversible, and an experimental measurement is available for R_1 (v_{exp}).

Another main advantage of stoichiometric models is that they require few parameters for simulation. Note that in the previously mentioned mass balances, no experimental

parameters are needed, and only a few experimental measurements are required for the flux inequality constraints. This is radically different for kinetic modeling, where each reaction mechanism is modeled with one or more kinetic parameters [17], which are not always readily available. Stoichiometric modeling stands then as a popular way to circumvent this unavailability of kinetic data.

Thanks to the development of whole-genome sequencing, stoichiometric models were able to become genome-scale in the early 2000s, i.e. covering most metabolic pathways in the cell. Since then, they are also referred to as genome-scale metabolic models, or genome-scale models (GEMs) for short. They are typically reconstructed in a semi-automatized manner, and account not only for metabolites and reactions, but also for genes and the corresponding gene-reaction relationships. Therefore, GEMs are perfectly suited for metabolic engineering, as they can be used to assess the effect of genetic perturbations in metabolism, such as knockouts or over-expressions of specific genes.

However, GEMs are by no means the best modeling approach and present several disadvantages compared to kinetic modeling [16]. As the number of reactions in GEMs is typically larger than the number of metabolites, these models almost always yield an underdetermined problem, i.e. the number of variables is lower than the number of linearly independent equations. This means that the number of solutions that satisfy the requirements is infinite (**Table 1**), and additional assumptions are needed to obtain a single solution, e.g. assume that the cell has a metabolic objective (**Section 1.3.2**). Furthermore, although the approach is useful for predicting the values of metabolic fluxes, it tells us nothing about metabolite concentrations and to what degree enzymes are saturated inside the cell, which would be relevant information for strain improvement in metabolic engineering. Therefore, alternative approaches have been proposed that reconcile kinetic data into stoichiometric models, either for only a handful of extracellular metabolites [18] or by only using simplified kinetic mechanisms [19].

1.3.2. Simulating genome-scale models: Wearing and tearing

Due to their linear structure, GEMs are highly efficient to compute: a single simulation of the network takes less than a second of computational time in a normal desktop computer. Given this efficiency, in almost 20 years of progress in the field since the first GEM was published [20] numerous constraint-based approaches for simulation of GEMs have been published [21]. All these approaches require the previously mentioned constraints of mass balance and flux capacity (**Section 1.3.1**), which create what is known as a "solution space", consisting of infinite possible flux distributions that satisfy these constraints. However, as we cannot study an infinite number of distributions, these approaches need additional assumptions to simplify the analysis.

Constraint-based approaches can be broadly classified into biased approaches, where an objective function is defined, and unbiased approaches, where there is no objective function [21]. Here, an objective function is defined as an (often linear) combination of the fluxes in the network set to be either minimized or maximized. This objective function should represent a goal that the cell experimentally shows to try to achieve, for instance maximizing its biomass yield or ATP turnover [22]. An overview of some constraint-based methods is presented in **Box 1**, where flux balance analysis (FBA) stands out as the first one implemented [23] and one of the most popular approaches to date. Most of these methods

come included in computational simulation toolboxes such as the COBRA toolbox [24,25] and the RAVEN toolbox (**Paper VII**).

Box 1: Constraint-based approaches used in this thesis to simulate flux distributions of GEMs.

- **Flux balance analysis (FBA):** A biased approach in which an optimization problem is solved to find a single flux distribution that minimizes/maximizes a defined linear function [26]. Note that this can yield equally optimal alternative solutions (**Figure 3**).
- **Parsimonious flux balance analysis (pFBA):** A 2-step optimization, where after a regular FBA simulation, the objective function is fixed and the absolute sum of all fluxes is minimized, to find the most "compact" solution [27].
- **Flux variability analysis (FVA):** An extension of FBA in which for each reaction of the network, the span that each flux can vary while preserving optimality is computed [28].
- **Random sampling:** An unbiased approach in which the interior of the solution space is sampled without imposing any objective function [29,30]. Combined with FBA, it can be used to sample only the optimality region (**Figure 3**).

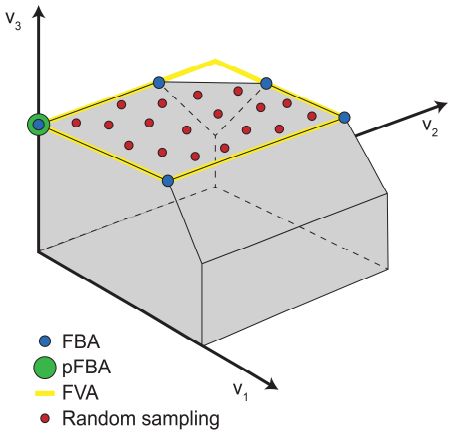

Figure 3: Results of different simulation approaches for a hypothetical solution space, using f(v) = v_3 as objective function to maximize [15]**.**

1.3.3. Evaluating quality of genome-scale models

For many organisms, multiple GEMs have been developed [31]; thus, when model developers create a new model, they need to compare it to previously existing models of the same organism, if available. Additionally, model users should have an easy way of knowing which model is better suited for their modeling objective. An even finer distinction is to compare among model *versions*, as models are often updated to newer versions whenever new biochemical knowledge or tools for improving model quality are published. Several metrics have therefore been developed for assessing model quality, both in terms of coverage, consistency of biochemical knowledge, and predictive power of simulations.

Typical evaluation metrics that asses coverage are the size of the model with and without accounting for compartmentalization (i.e. counting a metabolite only once if it repeats in different compartments) [32], the heterogeneity between models [33], and how well annotated the model is [34]. Regarding consistency, it is common to assess mass balance and charge balance of reactions in the model [35], network connectivity [36], and the existence of dead-end metabolites, i.e. metabolites that cannot be produced or consumed in the model, and/or blocked reactions, i.e. reactions in the model that cannot carry any flux. The last two metrics can be computed using FVA (**Box 1**).

Regarding predictive power, the most popular approach is to use FBA to predict the flux distribution under a given experimental condition. To validate predictions, intracellular fluxes can be quantified using a technique known as ^{13}C-based flux analysis [37]. However, this technique is experimentally challenging and restricted to only a few selected pathways; much more common instead is to compare flux predictions to experimental rates that are easier to measure, such as i) consumption/production rates of chemical compounds, and ii)

cellular growth. The first group of fluxes is predicted by reactions present in GEMs known as exchange reactions (**Figure 2**), which are included to allow mass to enter and leave the modeled system. In turn, growth can be predicted by including in the GEM what is known as the biomass pseudo-reaction, i.e. a mathematical representation of cell growth that lumps all the biomass components into a "biomass" pseudo-metabolite, which is excreted from the system in the same fashion of exchange reactions. The stoichiometry of this biomass pseudo-reaction is based on the abundances of the main cellular components, which have to be experimentally measured, and the energy demand needed for growth, referred to as the growth-associated ATP maintenance (GAM), which is typically fitted together with its non–growth counterpart (NGAM), to have the model match experimental data [23].

Another common approach for assessing predicting performance of GEMs is the ability of the model to reproduce experimental gene deletions, which is typically achieved by blocking the reactions associated to each gene and observing if the model is able to predict growth [38]. Caveats when assessing this metric (such as modeling setup, growth thresholds and experimental data) are available in **Paper I**.

It is important that all previously introduced metrics are readily available for anyone to compute on any model. As part of my PhD studies, I have developed a toolbox that measures most of the abovementioned metrics for comparing yeast models, available at https://github.com/BenjaSanchez/yGEMe. However, this toolbox is rather organism-specific and cannot be easily implemented for any model. In this regard, recent efforts that compute a suit of metabolic model tests automatically for any model (**Paper VIII)** are a great advantage for new models being developed.

1.3.4. Integration of omics data in genome-scale models: The battle of evermore

As previously mentioned, an important disadvantage of GEMs is that their simulations are undetermined, i.e. an infinite number of solutions are attainable if we only use mass balances and flux capacity constraints. Therefore, numerous methodologies have been developed for further constraining the solution space by adding different types of so-called omics data, i.e. experimental data that measures complete layers of biomolecular information in a high-throughput way [39]. By integrating these data, GEMs become a combination of the "bottom-up" and "top-down" approaches in one framework. The most commonly used types of data for this purpose are transcriptomics, i.e. gene expression; interactomics, i.e. gene-gene, protein-gene and protein-protein interactions; proteomics, i.e. protein intracellular levels; thermodynamics, i.e. reactions' physicochemical properties; kinetics, i.e. operational mechanisms of reactions; metabolomics, i.e. intracellular metabolite levels; lipidomics, a type of metabolomics data only accounting for lipid species; and fluxomics, i.e. measurements of flux values inside the cell or between the cell and its environment. A detailed explanation of each type of omics data together with examples of integration into GEMs of yeast is available in **Paper I**.

In general, omics data can be integrated in either a "hard" approach or a "soft" approach. In hard approaches, the data is directly used as additional numeric constraints (either as equalities or inequalities) on the metabolic fluxes, whereas in soft approaches no direct constraints are used and instead the model is simulated either with an objective function designed to come closer to the omics data levels, or *as is* to be later compared to the omics data [15]. In this thesis, I will mainly use the hard approach in my analysis, by using absolute

abundance of some biomolecules as constraints on metabolism. I call these "abundance" constraints.

There are currently many challenges when it comes to integrating omics data into GEMs. An important challenge, addressed further on in this thesis (**Chapter 4**), is the integration of proteomics, the measurement of every single intracellular level of protein inside the cell. This type of omics data has become in the past years more and more popular, especially with the development of mass spectrometry (MS) as a genome-wide approach for quantifying absolute protein copy number [40]. However, compared to the rest of the omics data types, it is the one that has been the least used in combination with GEMs (**Paper I**). Another challenge is the integration of several layers of omics data at the same time, which presents a set of challenges of its own regarding data consistency, which will be highlighted throughout this thesis.

1.4. Genome-scale models of yeast: Fifteen years in the light

The first GEM of yeast was published in 2003 [41]. Not only was it the first GEM for *S. cerevisiae*, but also the first eukaryal GEM; it comprised of 1175 reactions, 584 metabolites, 708 genes and 3 different cellular compartments: cytosol, mitochondria and extracellular space. Since then, 13 additional models have been released using this model as original template (**Figure 4A**). Each of these models has either provided additional simulation capabilities and/or improved coverage [42,43], mostly in terms of reactions and metabolites (**Figure 4B**, **Table 2**). Among these models, the consensus genome-scale network reconstruction project (formerly *yeastnet*, currently *yeast-GEM*) deserves a special mention, as it was created from the merge of two different models and manually curated using a 'jamboree' approach where several yeast research groups worked together at a three-day event [44]. Afterwards there has been several new published versions [45–49] and it is to date the main simulation-ready knowledge base of yeast metabolism.

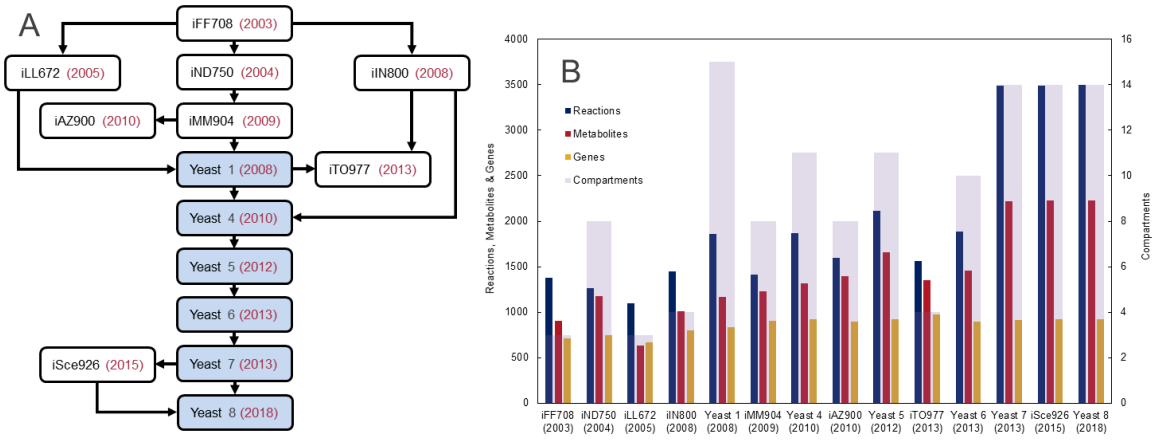

Figure 4: Genome-scale models of yeast. (A) Schematic history of GEMs in yeast. The models that are part of the yeast consensus GEM project are highlighted in light blue. (B) Number of reactions, metabolites, genes and compartments in all GEMs of yeast.

Table 2: Additional details of all GEMs published of yeast.

Name	Year	Novelty / additional details	Reference
iFF708	2003	First GEM of *S. cerevisiae*.	[41]
iND750	2004	Increased the number of compartments to eight.	[50]
iLL672	2005	Tested gene essentiality predictions under five environmental conditions.	[51]
iIN800	2008	More detailed lipid metabolism.	[52]
Yeast 1	2008	First consensus genome-scale network reconstruction.	[44]
iMM904	2009	Improved gene essentiality predictions. Integrated metabolomics data.	[53]
Yeast 4	2010	More detailed lipid metabolism. Allowed constraint-based simulations.	[45]
iAZ900	2010	Reconciled growth prediction inconsistencies.	[54]
Yeast 5	2012	Improved sphingolipid metabolism.	[46]
iTO977	2013	Decreased number of compartments and increased unique reactions.	[32]
Yeast 6	2013	Refined coverage and improved anaerobic predictions.	[47]
Yeast 7	2013	Enhanced fatty acid, glycerolipid and glycerophospholipid metabolism.	[48]
iSce926	2015	Improved Gene Essentiality and Synthetic Lethality predictions.	[55]
Yeast 8	2018	First yeast model tracked using a version control system.	[49]

Previous studies have compared some of these models in terms of prediction capabilities. One study showed that models of yeast that have been manually curated are better in predicting intracellular fluxes when compared to experimental measurements from ^{13}C-based flux analysis [56], highlighting the value of manual curation in the development of GEMs. Another study showed that when comparing the similarities between models, they tended to cluster based on the research group that had developed them, and that no model was better than the rest when predicting gene essentiality. Here, it is important to note that the biomass pseudo-reaction has remained almost entirely the same for all yeast models [57], suggesting that a potential improvement for prediction performance could be to account for more biomass components, using more detailed abundance data.

The previously introduced GEMs of yeast have been used in numerous occasions for metabolic engineering applications, such as for assessing mutant phenotypes [58], increasing production yields [59], and assessing optimal co-cultivation strategies with other species [60]. These and more applications are reviewed elsewhere [16]. With the development of new experimental and computational methods for improving model quality, I expect the number of applications to continue increasing in the future.

1.5. Aims and significance

Until here, I have introduced the genome-scale modeling approach, and how it has been applied to better understand yeast physiology (**Paper I**). In particular, I introduced the concept of "abundance constraints" as constraints that are defined by omics data containing absolute quantities of biomolecules. These constraints could be highly useful, as they represent physical limitations inside the cell, as opposed to *ad-hoc* constraints. This thesis will explore abundance constraints in yeast metabolism, by connecting GEMs to two different levels of omics data: lipidomics (measurements of lipids) and proteomics (measurements of proteins, particularly enzymes).

The thesis is divided in three parts (**Figure 5**). In the first part (**Chapter 2**), I lay two foundation stones to properly address the integration of omics data in GEMs. The first foundation stone is traceability of GEM development, which is needed to guarantee that the performed analysis can be reproduced by others. For this, I introduce a version control strategy for recording changes in GEMs (**Section 2.1**). The second foundation stone is consistency of omics data, which is fundamental if we are to analyze several layers of information combined. For this, I present a dataset of *S. cerevisiae* grown under different levels of environmental stress: heat stress, osmotic stress and ethanol stress (**Paper II**). Understanding and properly modeling metabolism is essential for comprehending the stress response in *S. cerevisiae*, as stress causes increased energy demand and reorganization of the biomass composition.

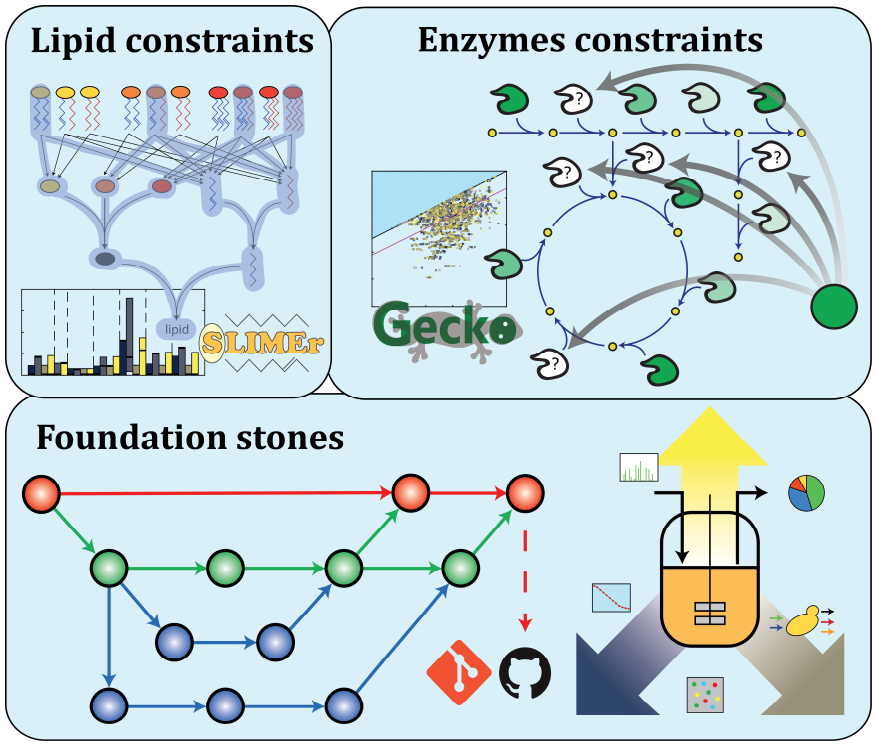

Figure 5: Graphical abstract of the research presented in this thesis. Top left block: The addition of lipid constraints through SLIMEr enables a flexible way for the model to satisfy lipid requirements. Top right block: The addition of enzyme constraints through GECKO allows to compute enzyme usage of every enzyme in the model at varying experimental conditions. Bottom block: Traceability model development through version control (left side) and consistent data integration through multi-omics studies (right side) are foundation stones for computing abundance constraints in metabolism.

Having set these foundation stones for proper data integration, In the second part of the thesis (**Chapter 3**) I investigate lipid metabolism. Here, a recurrent problem facing GEMs is to correctly represent lipids as biomass requirements, due to numerous combinations of individual lipid species and the lack of fully detailed data. In this thesis I present SLIMEr (**Paper III**), a formalism for correctly representing lipid requirements in GEMs using commonly available experimental data. I implement this approach in *yeast-GEM*, to make it possible to explore the flexibility of lipid metabolism at varying experimental conditions.

In the final part of the thesis (**Chapter 4**), I turn to enzymes and their relationship to metabolism. First, as GEMs do not account for enzymatic information, I present GECKO (**Paper IV**), a method for including enzyme constraints in GEMs based on kinetic and proteomics data. These enzyme constraints are implemented in *yeast-GEM* to investigate physiological behavior that could not be explained using regular genome-scale modeling. Secondly, as notorious variability is observed in the proteomics data measured in **Paper II**, I asses the protein quantification technique used to generate said data. I achieve this by introducing a separate dataset that focuses on biological and batch replicability (**Paper V**), and by measuring in this dataset accuracy and precision of different variants of the protein quantification methodology. Finally, I apply the GECKO formalism on the stress dataset, to create enzyme-constrained models for each of the experimental conditions, with flux limitations based on the proteomics detected values (**Paper VI**). These models are then analyzed to find trends in enzyme usage between and within stress types.

The analysis presented in this thesis is an example of how integration of mathematical modeling and omics data can yield novel insight into cellular physiology.

2. Foundation stones for the next generation of genome-scale models

As mentioned earlier, the field of genome-scale modeling has been moving more and more towards the integration of multiple levels of omics data, to infer previously unknown physiological behavior [61]. Particularly, this thesis will explore methods for connecting GEMs to two different levels of omics data: lipidomics and proteomics. However, it is relevant to set some foundation stones before proceeding with any further analysis. In this chapter I go through two main concepts that I see as the basis for robust genome-scale modeling: traceability of the model development, and consistency of the data used. Concerning traceability, I present a sustainable way of developing genome-scale models which uses version control tools as a way of keeping track of every change in a model, to guarantee that no unexpected changes are introduced that could decrease model quality. Concerning consistency, I present a multi-omics dataset of yeast grown under different levels of stress (**Paper II**) that will be used throughout this thesis, to guarantee that the different types of data we integrate into the model are consistent amongst themselves.

2.1. Sustainable development of genome-scale models: No surprises

An important challenge in genome-scale modeling is to properly evaluate GEMs, to determine if the changes we have done to a model are improving coverage and/or prediction performance (**Section 1.3.3**). However, a challenge that comes with this is to properly keep track of the changes as we develop a given model, so that we can guarantee reproducibility of our research [62]. In this section, I go through a new way of recording changes and developing GEMs that some models in my research group have started to follow. This section is not included in any of the papers that are part of my thesis, although the concept is briefly introduced in **Paper VII**, and the implementation of these ideas in the consensus GEM of yeast is part of **Paper X**.

Due to the size of metabolic networks, it is challenging to keep track of changes as we develop a GEM, making comparisons among different versions of the model difficult (**Figure 6A**). When only one researcher develops the model this can perhaps be overcome by ensuring that every change is manually recorded in some log; however, this becomes increasingly harder if two or more researchers develop the model together, as different tasks might overlap in the same components of the model, e.g. correcting stoichiometry of reactions and including reaction annotation might end up changing the same reaction, generating a conflict which is hard to resolve later. Additionally, when a group of researchers are developing a model it would also be beneficiary to have an on-line alternative for model

developers to quickly share changes to the model, and for users to obtain the latest version of the model and contact the developers if any problems with the model are detected.

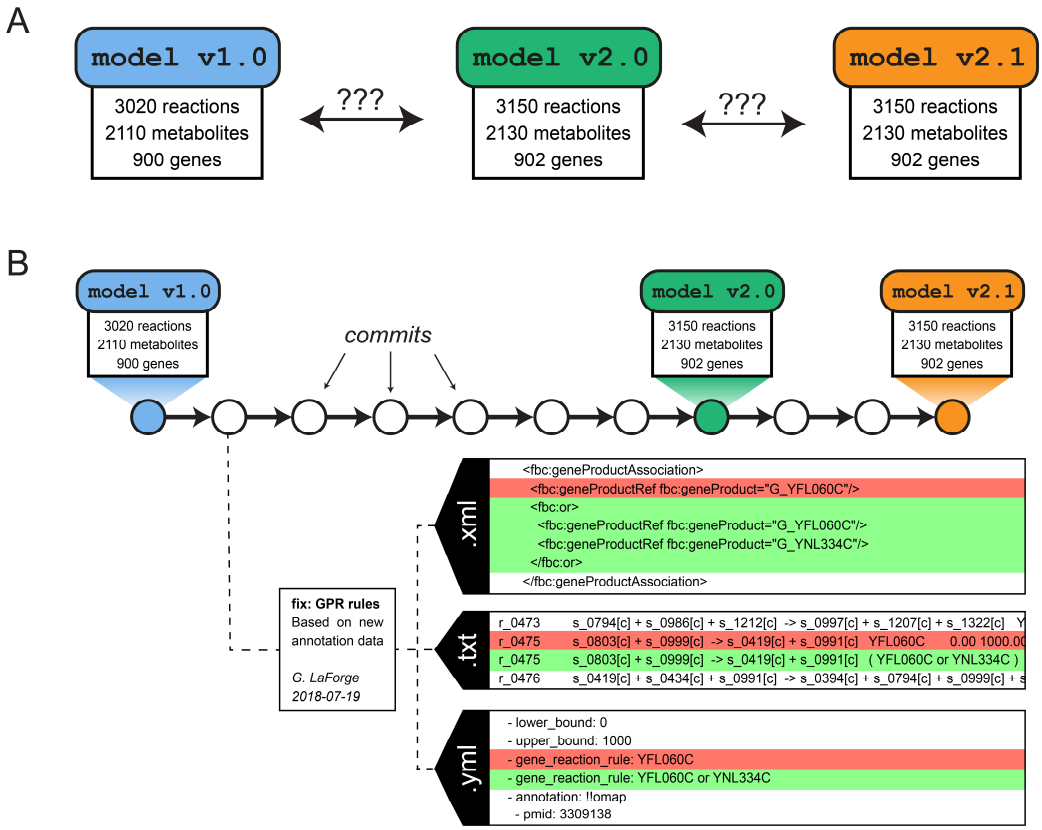

Figure 6: Different strategies for developing a GEM. (A) By editing the model without any version control system, it is hard to assess what are the differences between two versions of the model, as even if they would be of the same size (models 2.0 and 2.1), they could have differences. (B) Thanks to version control, every change to the model is registered in a *commit*, showing what was the change, who did it and when.

In software development, these problems have been solved with version control, i.e. the practice of automatically tracking changes to a file or set of files over time [63]. When using a version control system, a group of files is organized in a so-called "repository", and changes performed to the files are grouped in "commits", which later can be queried individually to understand the changes that are being introduced to the files. Git (https://git-scm.com/) is currently the most common alternative for version control, due to its efficiency and scalability. Additionally, web services such as GitHub (https://github.com/) and GitLab (https://gitlab.com/) offer free on-line hosting of Git repositories, for people to collaboratively develop code. Version control practices have become the standard to guarantee reproducibility in most scientific fields that uses computational analysis; it is then not hard to imagine that such a system could be implemented for tracking GEMs as well (**Figure 6B**), as GEMs are typically stored in file formats compatible with version control.

Even though several software for automatic GEM reconstruction exist [64–66], they are not designed for model storage and therefore do not account for any type of version control. Some alternatives for hosting models with version control do exist [67,68], but they rely on dedicated software (limiting the capabilities of model development), and more importantly do not provide an easy way of delivering the model to users and to allow the engagement of

the community. We therefore need a simple tool that can keep track of changes, allow work in parallel, minimize conflicts (if they appear), publicly store different versions of the model with a detailed log of every change, and allow users a straightforward way for contacting the developement team for issues and/or sending their own contributions.

Here, I outline a strategy for storing and developing a GEM with version control tools to promote reproducibility and open collaboration. The strategy has been implemented in *yeast-GEM*, the continuation project of the consensus GEM of *S. cerevisiae* [44]. In this strategy, Git is used as version control tool to track changes of the model files, and GitHub as hosting service to provide all files to the community. The model can be modified locally using Matlab® and saved using a RAVEN/COBRA wrapper function that stores the model in three different formats: the interchange systems biology markup language (SBML) format (*.xml*), meant for simulations across programming languages and toolboxes, and two summarized text files (*.txt* and *.yml*), meant for easier visualization of changes in GitHub or any Git graphical user interface (GUI) client (**Figure 6B**). The researcher modifying the model should also provide the corresponding data/scripts used for modifying the model, and use semantic commit messages [69] to describe what has been done in the model. This way, by looking at the history of the repository, anyone can know what has been changed, who did it and why, achieving a 100% traceable history of the model development.

To allow multiple developers to work in parallel in the model, and users to submit their own contributions, the strategy also entails a branching model. In version control, a "branch" refers to a copy of the files in the repository that can be modified with extra commits without modifying the rest of the project. In our strategy for GEM development, three different types of branches are accounted for [70] (**Figure 7**): a single "master" branch, which only gets updated with official new versions of the model; a single "development" branch, where all of the finalized work by the development team is kept; and several "feature" branches, which have the work in progress of each developer of the team (although one developer might have more than one branch). Both master and development branches can only be modified by an administrator of the repository.

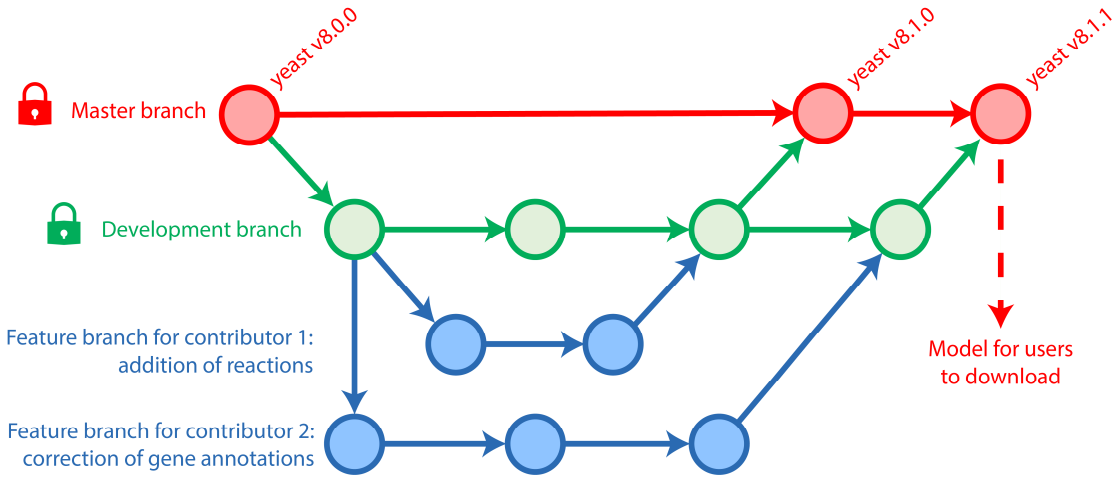

Figure 7: Diagram exemplifying how the consensus GEM of yeast allows for work in parallel by multiple collaborators. Different colors indicate the three different types of branches the repository contains. The lock icons indicate the branches that only the administrator can access. This scheme is a simplification of a popular branching model [70].

With this system, each researcher can work separately on their own branch, without interfering with the work of others. Once they have finished their specific project, they can request the administrator to merge their changes into the development branch, by opening what is known as a "pull request", which needs to first be reviewed and accepted by someone else in the development team. Any conflicts are easily spotted at this point and must be resolved before merging. After this reviewing process, the administrator of the repository proceeds to integrate the changes into the development branch. Once enough work has been accumulated to the development branch, the administrator proceeds to release a new version of the model, by merging the changes in the development branch to the master branch.

Additionally, developing GEMs using this approach has the advantages of other software hosted in GitHub: Users can open issues if they detect errors in the model and/or missing biochemical information, and they can comment on existing issues, providing their expertise. Developers can display the work they are currently doing by organizing it in projects, to let users know if a certain feature in the model will soon be implemented. Finally, the administrator can tag specific commits (**Figure 7**) as stable releases, so users can utilize those versions of the model for simulation purposes. These releases also include the model in Matlab® (*.mat*) and Excel® (*.xlsx*) formats, meant for quick simulation and data-navigation, respectively.

In my group, we have been using this system for development of GEMs for the past 2 years. In the case of *yeast-GEM*, it has been beneficial as a tool for working in parallel and reviewing each other's contributions. When a problem in the model performance has been detected, it has been quite easy to trace the source of the problem and fix it. Finally, it has increased collaboration between model developers, and allowed engagement of the yeast community in the project. Several modifications done to the model throughout this thesis have been included via pull requests into *yeast-GEM*, the main example being the addition of SLIME reactions (**Chapter 3**).

In conclusion, version control is not only essential for reproducibility when developing code in research, but can also be adapted to develop GEMs in a transparent and reproducible fashion, minimizing conflicts when working in parallel, and engaging the user community. The implementation of this version control system for *yeast-GEM*, including all releases plus detailed guidelines for users, developers and administrators, is available at https://github.com/SysBioChalmers/yeast-GEM.

2.2. Multi-omics dataset of *S. cerevisiae*

As shown in **Section 1.3.4**, integrating omics data in GEMs is a common practice to further constrain the solution space and improve metabolic predictions. As a plethora of published experimental data is available, the most common way researchers do this is by using different published datasets. However, a challenge emerges when several layers of data are used together, as most often these layers have been measured in separate experiments, of which each might have been conducted with a different strain of the same organism, or a different experimental setup. As these variables can exert an important influence on the measured data, the corresponding layers of data might have inconsistencies between them, resulting in unfeasible simulations when integrated in a GEM. Therefore, studies that measure several layers of information at the same time [71] are valuable resources for consistent omics integration into GEMs. In this section I go through **Paper II**, wherein a dataset was generated

of *S. cerevisiae* grown under several conditions of stress. As this thesis focuses on the integration of data in GEMs for inferring novel observations, the focus of this section is on the data presented in **Paper II**.

Understanding the response of *S. cerevisiae* to different conditions of stress is of high importance for biotechnology applications [72], as doing so can enable the development of resistant strains that can endure harsher conditions and/or improve product yields [73]. Among the different types of stresses yeast can face, some of the most commonly studied ones are heat stress [74], ethanol stress [75] and osmotic stress (by increasing the concentration of salt) [76]. Considering the modeling tools introduced in **Section 1.3**, it is valuable to gather omics data on the adaptation of yeast to these stresses, to look into the physiological responses from a systems biology perspective, and particularly to see how metabolism adapts to cope with said stresses.

To measure these data, we need an appropriate experimental setup. There are three main bioreactor setups that are routinely used in industrial biotechnology: batch, fed-batch, and chemostat [77]. The simplest one is the batch setup, where the organism is inoculated into a perfectly stirred vessel containing growth media, so that the organism consumes the nutrients in the media and replicates, until the media becomes limiting, i.e. runs out of one of the basic elements needed for growth, e.g. carbon. This setup allows measuring the dynamic response of the organism to variable substrate concentrations. The fed-batch setup starts in the same way as a batch setup, only that as soon as the media has become limiting, a feed of fresh media is slowly added to the vessel, to maintain low levels of the limiting element while still allowing cells to grow. This is especially beneficial if we wish to achieve high levels of biomass, as by keeping cells growing at a low growth rate we avoid production of secondary products that might inhibit growth. Finally, the chemostat setup (**Figure 8A**) is similar to the fed-batch setup, only that there is both a feed into the vessel and an outlet from the vessel, tuned so that the culture volume remains constant. By doing so, we can control the specific growth rate of the cells at the desired dilution rate of the vessel (the feed rate divided by the volume of media). In this study, the chemostat setup was used, as it is best suited for capturing the steady-state physiological response of cells, which will be later integrated with the genome-scale modeling approach.

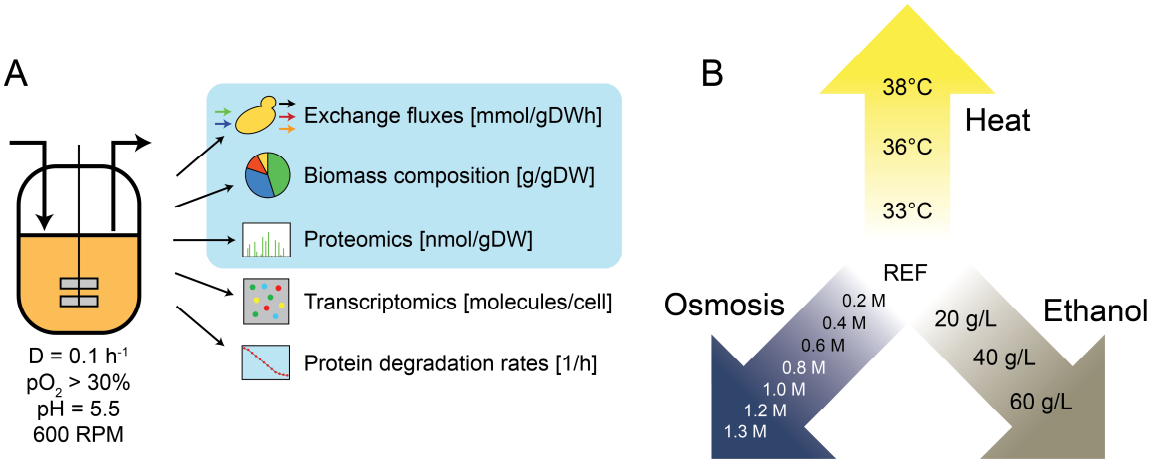

Figure 8: Experimental setup in this study. (A) The chemostat setup used to cultivate yeast, and all measurements performed on the extracted samples. The measurements used in this thesis are highlighted in light blue. (B) Summary of all studied stress conditions. The two different text colors (black/white) indicate the two different groups in which the MS data was measured.

Briefly, *S. cerevisiae*, strain CEN.PK113-7D, was grown in glucose-limited chemostats at a specific growth rate of 0.1 h^{-1}. The cultivations used minimal media, i.e. containing the minimum number of nutrients to allow growth, and were kept aerobic by sparging air so that the partial pressure of oxygen (pO_2) was always above 30%. Furthermore, a pH of 5.5 and a stirring rate of 600 RPM were kept constant for all chemostats. Using this setup, 14 conditions were assessed, with each one in triplicate (**Figure 8B**): a reference condition, grown at 30ºC and with no NaCl or ethanol; 3 levels of temperature stress; 7 levels of osmotic stress; and 3 levels of ethanol stress. Note that only 3 of the osmotic stress levels are included in **Paper II**, whereas the other 4 levels were separately generated for **Paper VI**.

For all the above-mentioned conditions, samples were collected from the steady state to analyze several layers of biological information (**Figure 8A**). The consumption/production rates of glucose, ethanol, acetate and other organic acids were inferred by measuring with high-performance liquid chromatography (HPLC) the concentrations of these substances in the outlet of the chemostat and performing a mass balance to deduce how much are cells consuming/producing. A similar analysis was done for the consumption of O_2 and production of CO_2, as the gases entering/leaving the system were also measured. All these rates can be compared to what the model would predict as exchange reaction fluxes [mmol/gDW] (**Paper VI**), or instead used as constraints on the simulations (**Paper III**).

Several established protocols exist for measuring the biomass composition, i.e. the mass fraction [g/gDW] that each component (protein, RNA, carbohydrate, etc.) take up inside the cell. Protein content was measured using a commercial assay kit, total RNA content using spectrophotometry, and trehalose and glycogen (two of the abundant carbohydrates in yeast) were separately isolated, hydrolyzed and measured with HPLC. Regarding lipids, ergosterol was detected with HPLC, and all other lipids were measured using both HPLC and an esterification process, as it will be explained in more detail in **Section 3.1**.

Absolute protein abundances [nmol/gDW] were also measured, by performing what is known as peptide-based shotgun MS [78], where the protein fraction is extracted, digested with a protease, and the resulting mix of peptides separated in a liquid chromatography (LC) column and detected using an MS instrument and a recognition software [79]. Afterwards, a technique known as intensity-based absolute quantification (iBAQ) [80] was used, which relies on an external commercial standard to infer the relationship between the peptide MS intensities and the corresponding protein abundances. This relationship was used to infer the protein abundances of a separate sample, consisting of proteins from a lysine auxotrophic strain of *S. cerevisiae* fed with heavy ^{15}N, ^{13}C-lysine. Finally, these values were used to infer the abundances of all the samples from the study with another approach known as stable isotope labeling by amino acids in cell culture (SILAC) [81]. When performing SILAC, we first mix in equal amounts each sample with the previously mentioned auxotrophic sample and then perform the MS analysis. This will yield, for each peptide (and therefore for each protein), two separate MS intensities, and as we know the abundance of one of them (thanks to the iBAQ technique), we can infer the abundance of the other.

Two additional layers of information are also accounted for in **Paper II**: transcriptomics data and degradation rates of proteins. Absolute levels of mRNA were inferred using next generation sequencing data from a previous study [82] combined with a commercial RNA quantitation assay to generate a standard curve that was applied on all transcripts and all experimental conditions. Degradation rates of proteins, referred to sometimes as protein turnovers, were inferred by using the previously mentioned SILAC technique to capture the

dynamic response of the proteome in the lysine auxotrophic strain of *S. cerevisiae* when transitioning from a feed with unlabeled to labelled lysine. Nonetheless, these two levels of information were not used in connection with genome-scale modeling in this thesis, so they will not be further analyzed here. The results for the other three types of measurements (**Figure 8A**) are summarized below.

It can be observed that exchange fluxes, the most direct way of assessing the metabolic response, undergo dramatic changes as stress levels increase. In the case of heat stress (**Figure 9A**), this only occurs at 38ºC, at which point yeast starts fermenting, i.e. producing ethanol. In the case of osmotic stress (**Figure 9B**), a change in the O_2 consumption and CO_2 production is only observed from 0.6 M, and production of ethanol only from 1.2 M. In fact, this observation led to measure additional levels of osmotic stress (0.8, 1.0, 1.2 and 1.3 M) for **Paper VI**. Finally, in the case of ethanol stress (**Figure 9C**), co-consumption of both glucose and ethanol were observed for all conditions, gradually increasing the uptake of glucose and decreasing the uptake of ethanol as the level of stress increased.

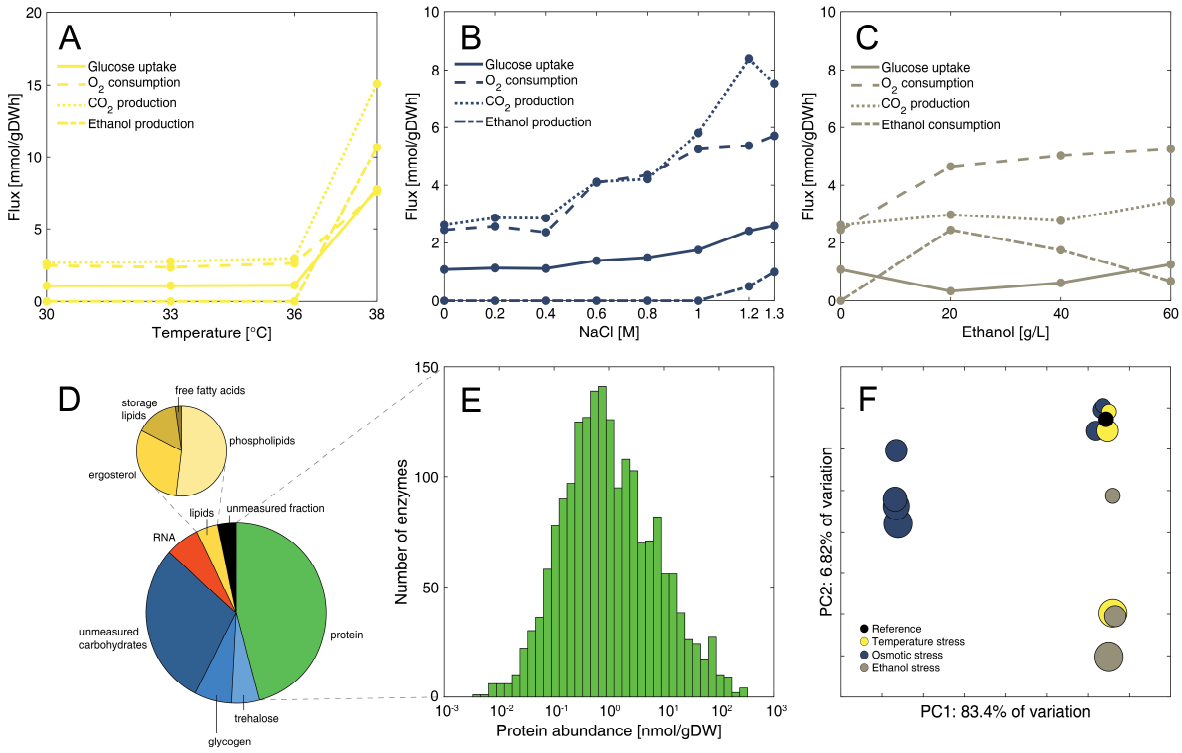

Figure 9: Overview of the experimental data. (A-B-C) Metabolic exchange rates at increasing levels of temperature (A), osmotic stress (B) and ethanol concentration (C). (D) Biomass composition at reference conditions. The undetected carbohydrate content is estimated from literature [41]. (E) Histogram of the protein abundances at reference conditions. (F) PCA of the proteomics data. Colors represent the stress types and marker sizes represent the stress level. The amount of variability that each component represents is also indicated.

Regarding the biomass composition (without accounting for water), it was seen that at reference conditions around 45% of the biomass corresponds to protein (**Figure 9D**); a fraction that increases further at higher levels of stress (**Paper II**). Note that from the carbohydrate fraction, which at reference conditions is known to be close to 40%, a large proportion was not measured, which is associated mainly to mannan and glucan [41]. For

the lipid fraction, which is close to 4% at reference conditions, around 80% of it corresponds to structural lipids (phospholipids and ergosterol), whereas the rest is primarly storage lipids (triglycerides and sterol esters).

Finally, regarding the proteomics data, around four orders of magnitude of variation were observed (**Figure 9E**) with a median protein abundance of 0.83 nmol/gDW, which is equivalent to ~6,500 molecules/cell, assuming an average cell mass of 13 pg. This protein distribution changes as the stress level increases, as can be inferred from a principal component analysis (PCA) (**Figure 9F**), a method that uses an orthogonal transformation to display in two dimensions all protein abundances under all conditions. Note here that most of the variation within the data is explained by the first component (x-axis), and that the two separated groups observed correspond to the separate times in which the proteomics data was measured (one for **Paper II** and one for **Paper VI**). The implications of this observation and different ways to address it are presented in **Chapter 4**.

3. Abundance constraints: Lipids

Lipids are essential to cells as the main membrane component; they have played a key role during evolution towards forming cells [83], as without them cells would not have become a self-contained unit. It is therefore essential to properly account for lipids when simulating metabolism. However, this poses a computational challenge, as there are many different combinations of lipids and most of them are hard to measure directly. In this chapter I go through **Paper III,** in which SLIMEr, a method for including typical lipid data in a GEM, was developed and implemented in the consensus GEM of yeast. The implementation of SLIMEr lead to an improved representation of lipid requirements, and allowed analysis of the high flexibility of lipid metabolism.

3.1. The challenge of integrating lipid data in genome-scale models

As previously introduced, simulation of GEMs relies heavily on the definition of a biomass pseudo-reaction [57,84], which consists of the abundances of the cell's building blocks. In the case of proteins, RNA and DNA, abundances are straightforward to obtain experimentally, as they are polymers built out of combinations of 20 amino acids, 4 ribonucleotides and 4 deoxyribonucleotides, respectively. Therefore, it is enough to isolate the corresponding fraction and detect the proportion of each unit, which can be achieved for amino acids using HPLC and for DNA and RNA using nucleotide sequencing [85]. Abundances are also relatively simple to obtain for carbohydrates, as while many types of carbohydrates are found in nature [86], specific organisms tend to contain only a few; in the case of yeast, most of its carbohydrate content are four different polysaccharides (glycogen, trehalose, mannan and glucan), which can be measured using standard protocols [87].

Obtaining lipid abundances is, however, more challenging. Most lipids are characterized by one or more nonpolar acyl chains, sometimes referred to as "tails", with varying length and number of saturations, connected by a smaller molecule referred to as "backbone", which defines the class each lipid belongs to. In yeast, over 20 different lipid classes and around 10 different acyl chains are normally expressed (some common examples are illustrated in **Figure 10**). As many enzymes that produce these lipids are not specific to particular acyl chains, practically any combination of backbone and acyl chains can be produced, yielding over 1000 different lipids yeast theoretically could express, out of which close to 250 have been experimentally detected [88]. Measuring each single lipid species is therefore challenging, and hence very few GEMs account for a biomass pseudo-reaction that has the abundances of all lipids [89].

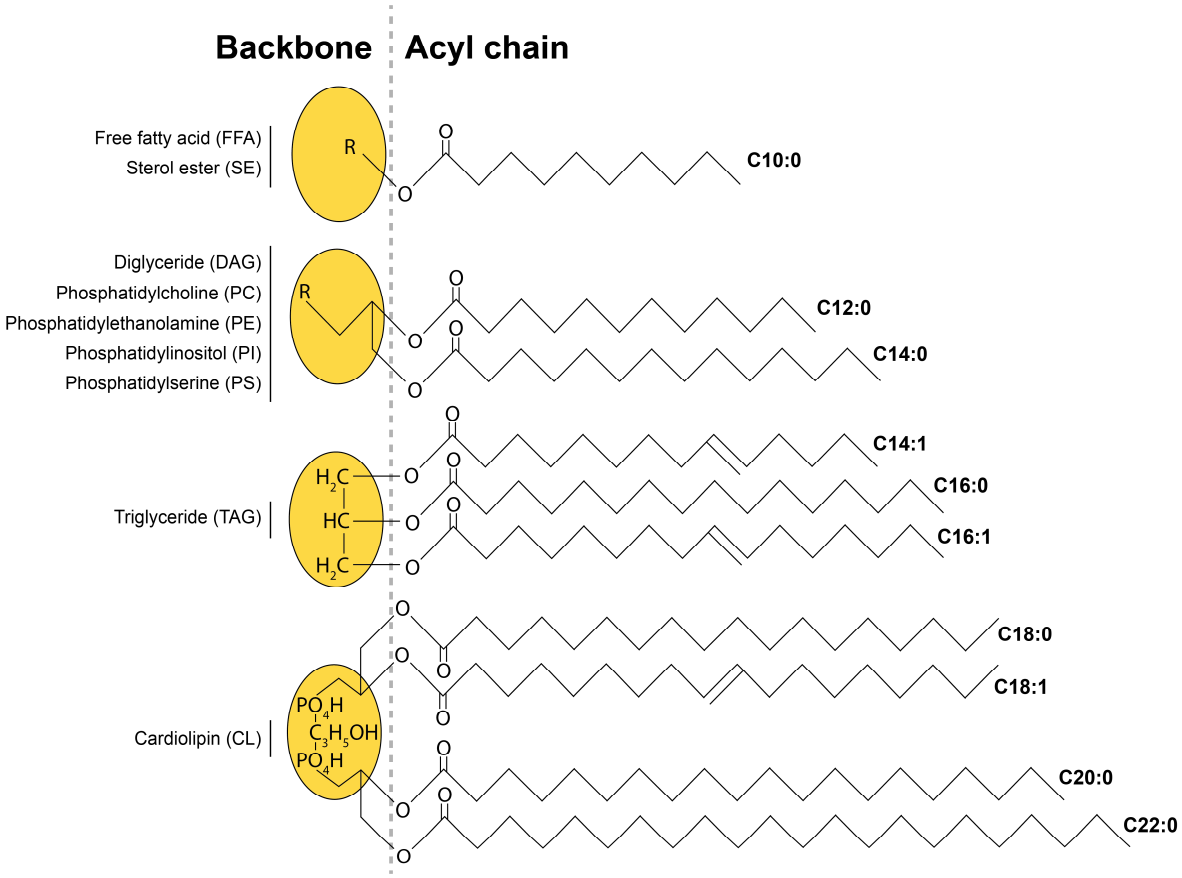

Figure 10: Some of the different types of lipid backbones and acyl chains that are commonly observed in yeast. Note that almost any possible combination of backbone and acyl chain(s) is theoretically possible.

A much more common approach for measuring lipid data is to independently measure abundances of each lipid class and each acyl chain. The former can be achieved with what is referred to as lipid profiling, which uses HPLC for separating the lipid classes (as lipids within the same lipid class will elute with a similar retention time in a chromatography column) [90,91], and the latter by an esterification process known as fatty acid methyl ester (FAME) analysis, which separates all acyl chains from the corresponding lipids, in connection with a gas chromatography (GC) for separating each individual acyl chain [92,93]. Therefore, GEMs have been adapted to depend on these types of lipid data, using either a "restrictive" approach or a "permissive" approach.

In the restrictive approach, a fixed acyl chain distribution is enforced on all lipid species, by creating a generic acyl chain pseudo-metabolite out of a combination of all acyl chains in a ratio that is based on the FAME data, which will then be the only acyl chain available for constructing any lipid species [52,94]. These generic lipid species are afterwards pooled together with a lipid pseudo-reaction, in a ratio that is based on the lipid profile data. The main disadvantage of this approach is that all lipid classes will be fixed to follow the same acyl chain distribution, which is known to vary significantly among lipid classes at different experimental conditions [95,96].

In the permissive approach, a lipid pseudo-reaction is also defined using the lipid profile data, but the generic lipids are allowed to be created from *any* specific lipid species [46,97]. The main disadvantage of this approach is that FAME data is disregarded, so the model is

free to choose any acyl chain for building up its lipids, which in practice will lead to lipids consisting mainly of short acyl chains, as those are less costly to produce in terms of energy requirements. These simulations would then be biased if longer acyl chains are experimentally detected. Consequentially, there is a need for an approach that can use both lipid class and acyl chain data, without enforcing a predetermined distribution.

3.2. SLIMEr: Split and conquer

To solve the above-mentioned challenges, we need to impose a separate constraint to each of the two "measurable entities" presented, i.e. the lipid backbones and the lipid acyl chains. However, as these entities are bound to the same metabolite in the model, we first need to mathematically represent the separation of both entities. Here I present SLIMEr, a method that adds reactions which Split Lipids Into Measurable Entities (SLIME); for each lipid species in the model, SLIMEr will add a pseudo-reaction that splits the molecule into its backbone and its different acyl chains. The stoichiometry of these reactions is proportional to the molecular weights of the associated entities (**Figure 11**), to convert molar flux [mmol/gDWh] into mass flux [g/gDWh], as the measured abundances typically come in mass units.

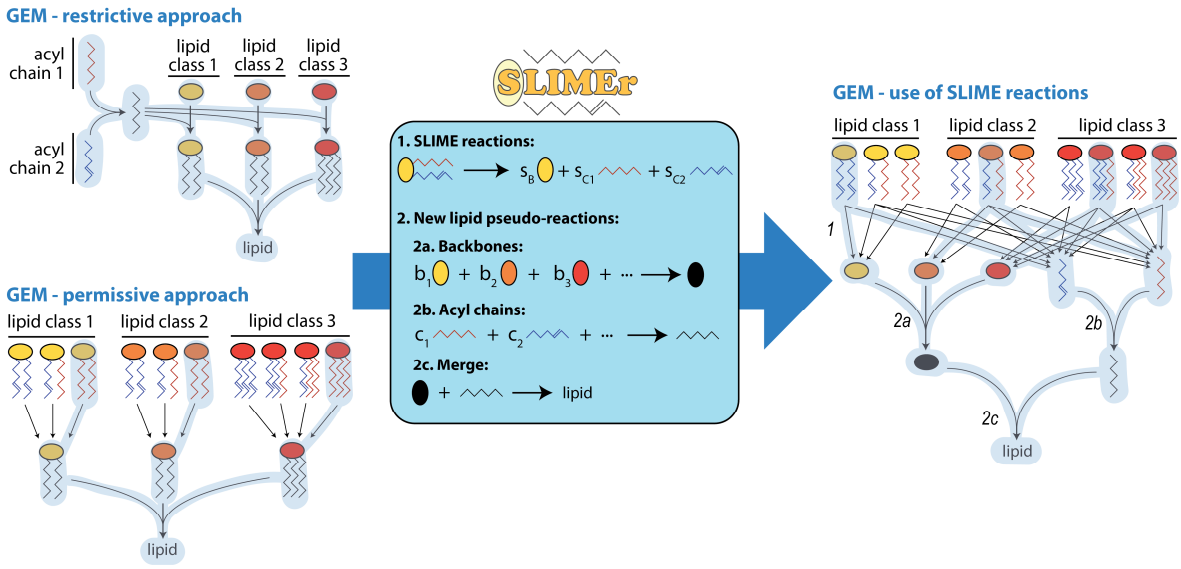

Figure 11: The SLIMEr formalism to improve lipid representation in GEMs. The active fluxes for a hypothetical flux simulation are highlighted in light blue. Top left corner: A restrictive approach for lipid metabolism results in all lipid classes having the same acyl-chain distribution. Bottom left corner: A permissive approach results in all lipid classes using the least energy expensive acyl-chain. Middle: Pseudo-reactions added to the model by SLIMEr. Right side: SLIMEr allows the model to satisfy at the same time the lipid class and acyl-chain distribution.

As now backbone and acyl chains are separated for each lipid, each of the two groups can be constrained. For this, SLIMEr adds 3 lipid pseudo-reactions (**Figure 11**): the first one pools together all backbones into a generic backbone, using the abundance data from lipid profiling as stoichiometric coefficients. The second one pools all acyl chains into a generic acyl-chain, using the abundance data from FAME analysis as stoichiometric coefficients. Finally, the third pseudo-reaction merges the generic backbone and generic acyl chain into

a generic lipid, which in turn goes into the biomass pseudo-reaction. Note that for the first two pseudo-reactions, as the abundances [g/gDW] are used for the stoichiometry, the incoming mass flux [g/gDWh] is now converted into lipid turnover [1/h], compatible with the biomass specific growth rate. Additional details on the formalism of this method are available in **Paper III**.

SLIMEr is therefore a truly unbiased approach compared to the restrictive and permissive approaches; on one hand, it allows FAME data to be used as input to the model, which in the case of the permissive approach would be neglected, yielding predictions biased towards shorter acyl chains. On the other hand, it does not enforce data onto each lipid class, which in the case of the restrictive approach would also bias the lipid distribution, as different lipid classes exhibit substantially different experimental acyl chain distributions (**Figure S2** in **Paper III**).

3.3. Improvement of the yeast model with the addition of SLIME reactions

SLIMEr was implemented in the yeast consensus GEM presented in **Section 2.1**, which until then was using the permissive approach for representing lipid requirements. For constraining the new lipid pseudo-reactions, the lipid profile and FAME data from **Paper II** were used. To guarantee coherence between the different sources of data, the lipid class abundances were rescaled to add up to the equivalent amount detected by FAME analysis, and the biomass composition was scaled to add up to 1 g/gDW [98]. After these modifications, the model remained of a similar size, with only 17 additional metabolites and 24 additional reactions. In the following, I refer to the original model as the "permissive" model, as *yeast-GEM* used the permissive approach until this point, and the resulting model as the "enhanced" model.

Both models were simulated under the reference condition of the stress dataset, constraining all measured exchange fluxes and using the maximization of ATP maintenance as objective function with a pFBA approach [27]. By using SLIMEr, the enhanced model is forced to have the same acyl chain distribution as the experimental data, whereas a permissive model prefers mostly shorter acyl chains (**Figure 12A**). SLIMEr is hence appropriate if we wish to have a model that simulates an experimentally meaningful acyl chain distribution.

Additionally, an FVA [28] of each group of SLIME reactions illustrated that the enhanced model retains a high flexibility, i.e. it has a wide range of possible acyl distributions that it can choose from (**Figure 12B**). Therefore, SLIMEr does not restrict the model excessively and is well-suited to study different experimental conditions where the lipid distribution might change significantly. Furthermore, the model achieved this new lipid configuration by only spending an additional 0.4% of the ATP maintenance needed in the model for unknown processes, a value that remained relatively constant at increasing levels of stress (**Figure 3C** in **Paper III**). This implies that achieving proper lipid compositions does not take a significant amount of additional energy compared to all metabolic costs, and physiological predictions of the model will overall remain similar after the addition of SLIME reactions.

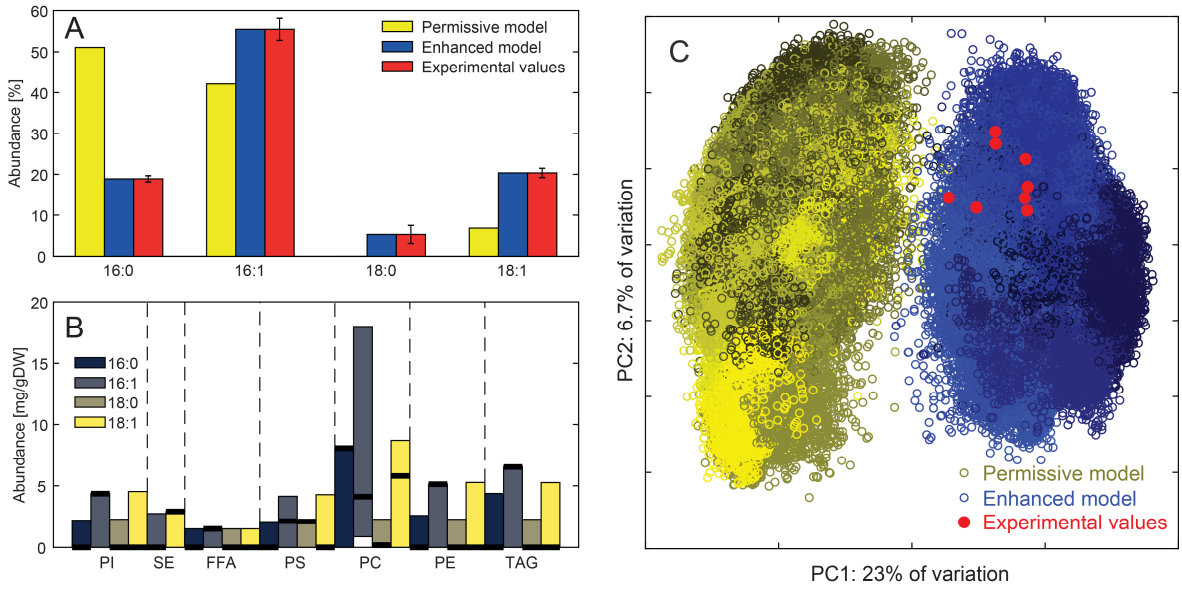

Figure 12: Implementation of SLIMEr in the consensus genome-scale model of yeast. (A) Acyl chain distribution at reference conditions for the permissive model (free to choose any acyl chain) and enhanced model (enforced to follow the experimental distribution). (B) Acyl chain breakdown predicted by the enhanced model, for each lipid class. Thick black lines are pFBA predictions and colored bars are the FVA allowed ranges. (C) PCA of lipid distributions experimentally measured and predicted by both models. 8 experimental conditions were assessed in total, and for each condition each model was simulated 10,000 times using random sampling and plotted with a different yellow/blue tonality, respectively.

I further tested the utility of SLIMEr by constructing condition-specific models for an additional dataset of 8 different experimental conditions, where 250 lipid species were measured [88], out of which 102 lipids (over 80% by mass) were present in the model. To construct the models, the lipid class abundances and acyl chain abundances were calculated from the measurements of specific lipid species. By performing random sampling [30] of both the permissive and enhanced models under all 8 experimental conditions, it is observed that the enhanced model is able to have a much closer lipid distribution to the original experimental values (**Figure 12C**). Nonetheless, the variability of the experimental data remains lower than the variability of model predictions; this suggests that even though there is a theoretically high level of flexibility in lipid metabolism, yeast has a high level of regulation in place to constrain its lipid composition depending on the environmental conditions [95].

In conclusion, by implementing SLIMEr the enhanced model i) enforces acyl chain requirements without decreasing the network flexibility nor significantly increasing the metabolic energy demand, ii) better predicts lipid abundance distributions, and iii) can compute the lipid requirements of transitioning between conditions. SLIMEr is available at https://github.com/SysBioChalmers/SLIMEr, and its implementation in the consensus GEM of *S. cerevisiae* is available from version 8.1.0 [99] onward.

4. Abundance constraints: Enzymes

Enzymes are cellular components that enable life as we know it, as they catalyze thousands of reactions inside the cell that otherwise would take much longer to occur or would not happen at all. As they are so closely related to metabolism, there has been a high interest in the past decade to integrate GEMs with enzyme information: mainly kinetic data, which tells us about the mechanics and operational rates of enzymes; but also proteomics data, which tells us about enzyme abundances. This chapter covers **Paper IV**, **Paper V** and **Paper VI**. **Paper IV** introduces a new method for integrating proteomics and kinetic data into GEMs and implements the approach in the consensus GEM of yeast. **Paper V** addresses the challenge of measuring proteomics data and proposes alternatives to increase the quality of said data. Finally, **Paper VI** uses the abovementioned method to analyze a large dataset of proteomics data to unveil previously unknown responses of yeast metabolism to stress.

4.1. Enzyme constraints in metabolism

Even though several computational approaches exist to connect metabolism to enzyme information, none of them allow for a genome-scale direct integration of a single proteomics dataset to create a condition-specific model. In this section I go through **Paper IV**, where GECKO, a method for adding enzyme constraints to a GEM, was developed and implemented in the consensus GEM of yeast. By doing so, I show that the enhanced model can predict physiological behavior that would otherwise have to be enforced with additional *ad-hoc* constraints. Furthermore, the enhanced model now allows to integrate proteomics data in a straightforward way and get insight into enzyme usage across metabolism.

4.1.1. Integrating metabolism and enzymes: Come together

GEMs have been thoroughly used as tools for predicting biological behavior [61]. A challenge in this regard is that GEMs most often rely on predetermined uptake rates of nutrients, as they have no internal "capacity constraints" to limit their intracellular fluxes with most fluxes in the network being completely unconstrained (**Figure 2**). These capacity constraints would be especially useful when studying biological conditions with high energy demand, such as during fast growth or under stress, as the cell is forced to process high amounts of substrate under these conditions, creating limitations in metabolism. In the case of GEMs, if we do not enforce an uptake rate, simulations yield either unlimited growth or unlimited tolerance to the stress, respectively. This is not observed in biology, as all organisms exhibit a maximum specific growth rate, and a maximum stress tolerance level. GEMs would therefore benefit from including capacity constraints.

Perhaps the most obvious capacity constraint is the limitation of enzyme levels, as enzymes catalyze most of the reactions occurring inside the cell. When an enzyme catalyzes a reaction, it momentarily binds to the corresponding substrate(s), which forces a conformational change that leads to the substrate(s) being transformed into the product(s), after which the enzyme releases the product(s) and is free to operate again. The time that it takes for the enzyme to complete a full catalysis cycle, together with the amount of enzyme available, imposes a constraint on every reaction catalyzed by an enzyme. In other words, the reaction's flux [mmol/gDWh] cannot exceed the enzyme's specific catalytic rate, often referred to as the turnover number or k_{cat} value [1/h], multiplied by the enzyme's intracellular concentration [mmol/gDW].

The previously mentioned limitation on metabolic reactions is valid for the case in which a single enzyme catalyzes a single reaction (**Figure 13A**); however, other cases are extremely common in the cell [100]. Some enzymes exhibit metabolic activity towards several different substrates (and hence reactions); they are therefore known as promiscuous enzymes (**Figure 13B**). It is also common to find several separate protein subunits that together form one catalytic unit, the latter known as an enzyme complex (**Figure 13C**). Finally, some reactions can be catalyzed by several different enzymes, which are referred to as isozymes (**Figure 13D**). Therefore, we need to take these complex relationships into account if we want to integrate enzyme information into GEMs.

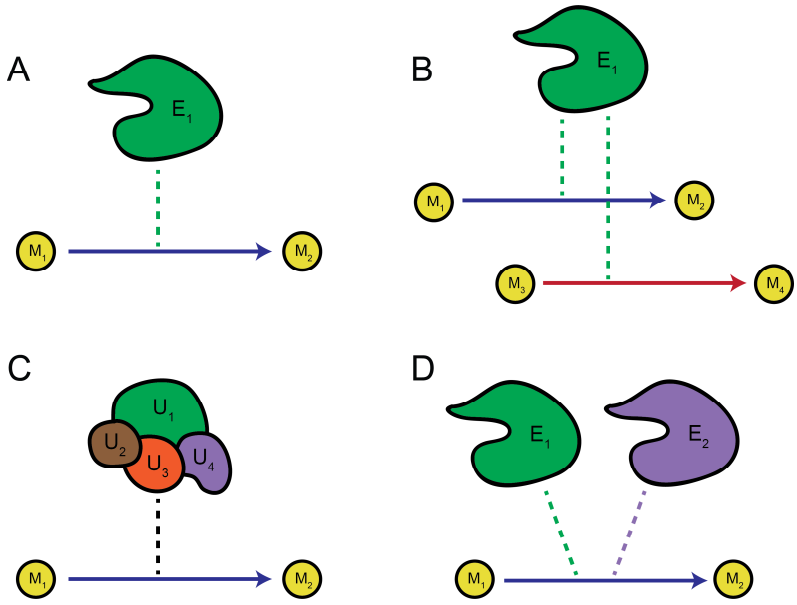

Figure 13: The different relationships between enzymes and metabolic reactions. (A) Only one enzyme catalyzes only one reaction. (B) A promiscuous enzyme catalyzing two different reactions. (C) An enzyme complex catalyzing a reaction. (D) Two isozymes that can each separately catalyze the same reaction.

Several genome-scale approaches have been developed to integrate the enzymatic and metabolic layers. One of the most popular approaches has been flux balance analysis with molecular crowding [19], which takes the idea of a limitation on enzymes' activities, and sums them all to impose a single global constraint on the volume that proteins take up inside the cell. This approach has been modified to instead constrain the mass fraction inside the cell [101], and has been implemented in *E. coli* [19,100,102], yeast [103,104] and cancerous human cells [101,105]. A main feature of this approach is that it does not rely on proteomics

data, but instead the model computes the amounts of each protein that the cell requires. This can however be a disadvantage if there are *in vivo* limitations for some specific pathways that do not come close to the optimal *in silico* usage distribution.

Some other approaches for combining metabolism with enzymes have also been developed, including modeling of RNA and protein synthesis processes starting from the transcription rates of genes [106,107], using proteomics datasets to infer hard constraints on fluxes [108], and even modeling most of the known processes in cell [109]. However, these approaches rely either on detailed biochemical knowledge of processes not fully understood in eukaryal cells, copious amounts of experimental data, and/or an excessive number of experimental parameters that are hard, or even impossible, to measure. Furthermore, something that none of these approaches allow for is directly integrating a single proteomics dataset on a genome-scale way, with a flexible manner to account for missing data.

4.1.2. GECKO: A simple tool for reducing complexity

To address the aforementioned challenges, I present a method for enhancing Genome-scale modeling with Enzyme Constraints, using Kinetics and Omics (GECKO). GECKO adds pseudo-metabolites and pseudo-reactions to the model, i.e. additional rows and columns, respectively, to the stoichiometric matrix. The pseudo-metabolites correspond to enzymes, which will act as substrates in their corresponding reactions. Their stoichiometries are defined as inversely proportional to their k_{cat} values (**Figure 14A**), to represent the fact that on a short timescale an enzyme has to be occupied momentarily and hence cannot be used twice at the same time. The pseudo-reactions in turn correspond to enzyme usages, i.e. the amount of enzyme occupied at a given moment by the corresponding metabolic task(s). This usage can be constrained with an upper bound equal to the measured abundance. With these additions, a mass balance for the enzymes yields the desired enzyme constraints we wanted to account for (**Figure 14A**). Note that this approach keeps the linear structure that characterizes GEMs, to in turn keep an efficient performance and compatibility with standard toolboxes [24,110].

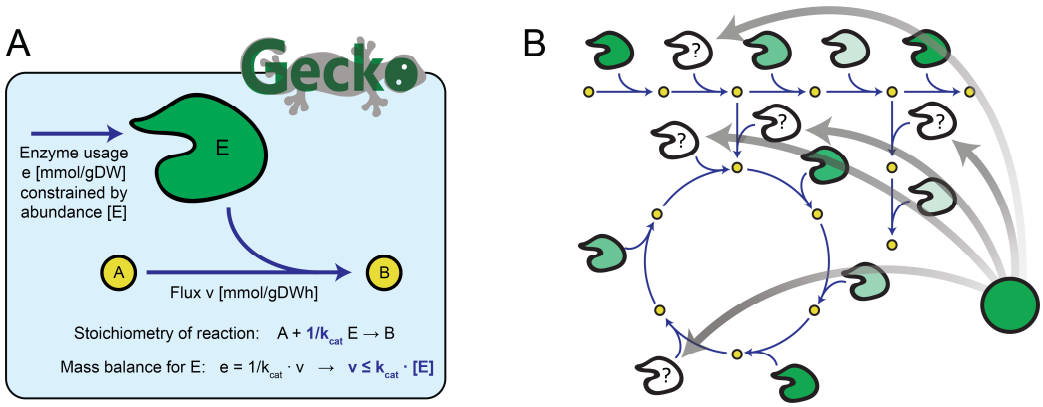

Figure 14: The GECKO formalism accounts for enzyme constraints in GEMs. (A) GECKO adds for each enzymatic reaction in metabolism the corresponding enzyme as part of the stoichiometry of the reaction, and an enzyme usage pseudo-reaction limited with the measured abundance of said enzyme. (B) For any enzyme for which an abundance was detected, GECKO adds a corresponding enzyme usage pseudo-reaction. For each undetected enzyme, GECKO connects that enzyme to a shared enzyme pool and then adds a single enzyme usage pseudo-reaction to said pool.

Most proteomics methods, including mass spectrometry, are not specific enough to detect every single protein, and rely on protein extraction steps that might not be able to separate some proteins. Therefore, many enzymes present in the model might not have any measured abundance in the proteomics dataset whilst still being present in the cell. Therefore, the model needs to cope with incomplete data without forcing some abundances to zero just because they are undetected. This is achieved by introducing a pseudo-metabolite that pools together all undetected enzymes, constraining only the mass of said pool (**Figure 14B**), using average abundance values from the literature [111] and an average saturation factor [104]. This allows the method to be flexible enough to receive partial proteomics data or even no proteomics data; for the latter case all enzymes will be connected to the abovementioned pool, in a similar fashion to the molecular crowding approach [19].

GECKO is designed to handle the previously introduced complex relationships between enzymes and reactions [112] (**Figure 13**). If an enzyme is promiscuous, the same enzyme usage will be shared by the different reactions it is associated to. If several subunits form an enzyme complex, then all subunits will be a part of the same reaction, and their stoichiometric coefficients will be adjusted by the corresponding subunit stoichiometry. If a reaction has several isozymes, then a separate reaction will be created for each isozyme, so that each isozyme can be used separately, and a so-called "arm" reaction will be created to keep the original upper bound of the reaction at the same value [113]. Finally, it is also relevant to note that the enzyme constraints described in this study only work for irreversible reactions (as in the opposite direction there would be production of enzyme). Therefore, GECKO splits reversible reactions in two separate irreversible reactions, adding the corresponding enzyme as substrate in both directions.

The kinetic data used in GECKO is retrieved from literature in a semi-automatic manner. First, all k_{cat} data from the BRENDA database [114] is downloaded, and for each reaction the best available k_{cat} value is queried: preferring first a measurement of an enzyme from *S. cerevisiae* and for the same substrate from the reaction, but otherwise prioritizing a measurement from the same substrate, then from the same organism, and then from the same enzyme class. Whenever more than one k_{cat} value is available, the highest one is chosen, to avoid over-constraining the model. After this automatic step, manual curation is performed based on previous data [104] and literature. Other enzyme information, such as gene-protein relationships and molecular weights of enzymes, is queried from additional databases, such as Swiss-Prot [115] and KEGG [116]. GECKO is publicly available at https://github.com/SysBioChalmers/GECKO and additional details can be found in **Paper IV** and its corresponding supplementary material.

Compared to the previously presented modeling approaches (**Section 4.1.1**), GECKO stands out as the most straightforward way for quickly creating a model that can i) account for enzyme constraints based on semi-curated kinetic data and ii) integrate proteomics data. Compared to the different molecular crowding formalisms [19,100–105], GECKO is better equipped for working at the genome-scale by accounting for non-ideal enzyme/reaction relationships, and it can incorporate experimentally determined protein levels from proteomics. Compared to other genome-scale approaches [106–109], it does not require extensive knowledge on protein synthesis and/or numerous datasets, as it only depends on knowing which enzymes catalyze which reactions, and it can run with proteomics data from a single experiment (or even with no proteomics data).

4.1.3. Improvement of the yeast model with the addition of enzyme constraints

By implementing GECKO on the yeast consensus GEM presented in **Section 2.1**, I created the enzyme-constrained model of yeast (hereafter referred to as *ecYeast*), which includes over 750 enzymes and over 3,200 reactions with enzyme constraints. The larger number of reactions is primarily because most enzymatic reactions are reversible, i.e. the method will split them in two reactions (see **Section 4.1.2**), and because over 300 enzymes exhibit promiscuity. Regarding k_{cat} values, 52% of the values came from *S. cerevisiae*, 31% matched the specific substrate, and 44% were of the same enzyme class. This coverage varied widely across metabolic pathways, with some pathways having no measurements available for yeast (**Figure 15**). Comparing *ecYeast* to the original GEM, both models exhibited similar topology and similar simulation running times on a standard growth maximization FBA problem using the COBRA toolbox [24]. However, they are widely different in terms of simulation capabilities, as I will summarize below.

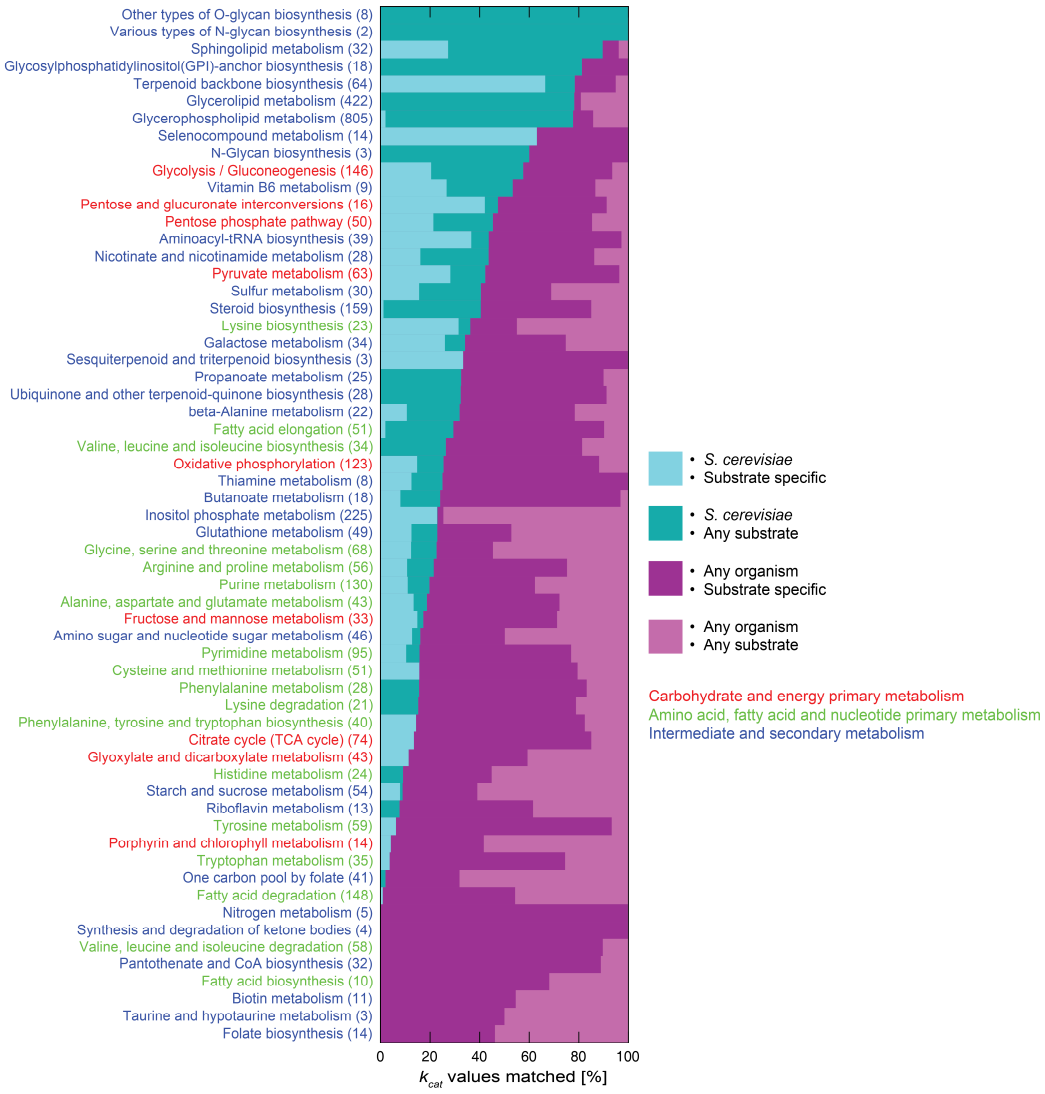

Figure 15: Coverage in *ecYeast* of k_{cat} values retrieved from the BRENDA database. Percentages are color coded depending on if the value came from measurements in *S. cerevisiae* and/or the exact substrate in the model. Pathways are color coded depending on the group of subsystems they belong to [117]. The number of enzymes in each pathway is shown in between brackets.

As previously stated, simulating maximization of biomass with a GEM relies on first defining a substrate uptake rate, as GEMs do not have any internal constraints to limit the amount of substrate that can be consumed. However, by accounting for enzyme constraints, there is now no need for pre-defining uptake rates. This was tested in simulations of *ecYeast* with no proteomics data, where it was observed that under several carbon sources for both minimal and complex media [118,119], this model came close to the measured maximum specific growth rates in aerobic batch cultivations (**Figure 16A**, average relative error of 8%) without needing to limit any substrate uptake rate. In turn, the original GEM would predict infinite growth for all conditions, and even if we would fix the uptake rates, predictions would still be overestimated (**Figure 4A** in **Paper IV**).

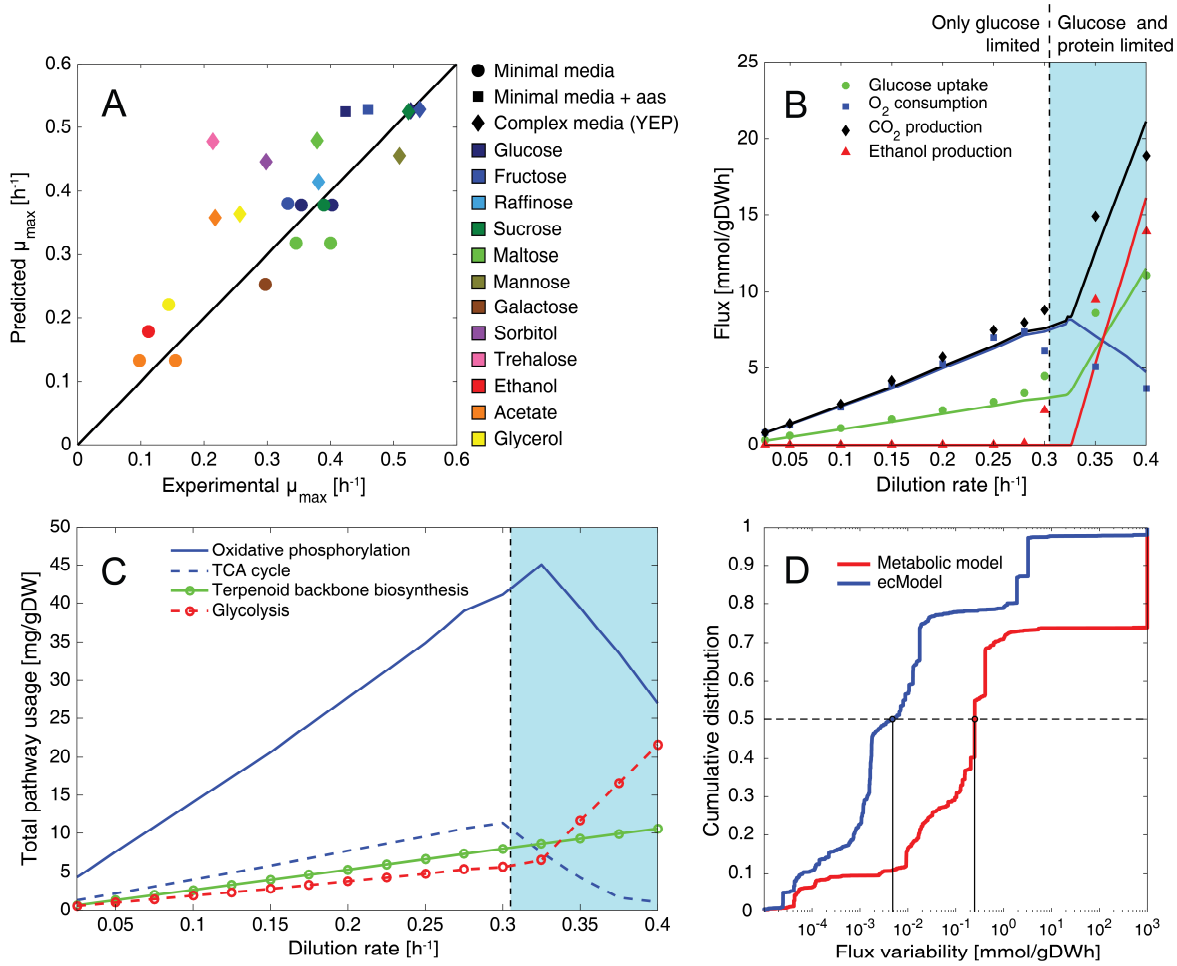

Figure 16: Implementation of GECKO in the consensus genome-scale model of yeast. (A) Maximum specific growth rate [h^{-1}] predictions of *ecYeast* versus experimental values from batch growth. (B) Exchange rates [mmol/gDWh] predictions of *ecYeast* versus experimental values from chemostat cultivations. (C) Summed pathway usage [mg/gDW] in *ecYeast* at increasing levels of growth. (D) Cumulative distribution of the variability span of all fluxes in the network, for the original yeast GEM and the *ecYeast* model constrained with proteomics data.

With *ecYeast* we can not only predict maximum specific growth rate, but also metabolic switches as a response of increased energy demand. *S. cerevisiae* is known to change its metabolic strategy for energy generation if it increases its growth rate over a certain threshold (the critical growth rate): from full respiration, the most energy efficient pathway in

metabolism, it switches to a mix of respiration and fermentation, the latter which produces ethanol. This phenomenon is known as the Crabtree effect [120] and is analogous to similar processes in *E. coli* [103] and cancerous cells [105]. There are numerous theories about what exactly drives this shift [121,122]; among them, protein allocation has been proposed as a possible explanation [103,104].

According to this theory, the switch occurs because even though respiration, and more specifically the oxidative phosphorylation pathway, is more efficient in terms of carbon, i.e. it has the highest ATP/carbon molar yield, it consists of enzymes that are large and not very fast, i.e. needs a large amount of protein to operate at higher rates. Fermentation is much more efficient in this regard, as it consists of smaller and faster enzymes, even though it yields less ATP per mole of carbon. As after the critical growth rate the cell enters a region referred to as the Janusian region [123] where it becomes both glucose-limited and protein-limited, a new strategy is needed to achieve high rates of ATP production: a progressive trade-off between respiration and fermentation.

I tested this theory in *ecYeast* (with no proteomics data), as this model accounts for the efficiencies of each enzyme, represented by their maximum specific rates. It is observed that the switch is indeed predicted (**Figure 16B**) following chemostat experimental data [124], whereas a regular GEM under the same conditions would instead show a linear increase in respiration with no metabolic switch, unless direct constraints on the exchange reactions would be used [125]. The metabolic switch can be further investigated through the total pathway usage, i.e. the sum of all enzyme usages in each specific pathway (**Figure 16C**): For anabolic pathways such as the terpenoid backbone biosynthesis (required for ergosterol production), there is no change in the enzyme usage slope after the critical growth rate, whereas there is a consistent decrease of both oxidative phosphorylation and the tricarboxylic acid (TCA) cycle, the latter needed for creating enough redox potential to enable respiration. In turn, the slope of glycolysis, and correspondingly fermentation, significantly increases, supporting the idea that fermentation gradually replaces respiration at high growth rates.

As a final study GECKO was tested as a way of reducing flux variability. Flux variability refers to the fact that, given the undetermined nature of GEMs, even though a specific solution is obtained when simulating the model with e.g. FBA, this solution is often non-unique (**Figure 3**), and many more solutions are equally optimal. In general, we wish to reduce this variability in our model predictions, as many of the possible solutions are not biologically meaningful. I tested whether flux variability could be reduced by performing FVA [28] under the reference conditions from the stress dataset (**Paper II**), on both the original model and the *ecYeast* model constrained with the proteomics data from the corresponding condition. Although the simulated exchange fluxes were similar on both models, the analysis showed that 64% of fluxes in the metabolic model reduced their variability after adding enzyme constraints, and overall the variability distributions were significantly different (**Figure 16D**). In particular, only 1.5% of the fluxes in *ecYeast* had maximum variability, compared to 25% of the fluxes in the original metabolic model. GECKO therefore proves to be a useful tool for decreasing variability of model predictions, while maintaining a physiologically relevant solution.

In conclusion, by implementing GECKO the enhanced model i) can predict maximum specific growth rates without needing to set any uptake rates, ii) prefers fermentation as an efficiency strategy at high growth rates, iii) allows studying enzyme/pathway usage at the

genome-scale, and iv) has reduced flux variability. The enzyme-constrained model of yeast is available at https://github.com/SysBioChalmers/ecModels.

Before moving on to the next section, it is important to address the fact that simulations of *ecYeast* rely to a large extent on the choice of k_{cat} values. For example, when predicting maximum specific growth rates under different carbon sources (**Figure 16A**), I assessed 10,000 simulations of the model but with random k_{cat} values based on a gamma distribution fitted to the k_{cat} data in the model. I saw that less than 10 of the simulations had a prediction error of 8% or less, whereas most models fitted the data poorly. I saw the same phenomenon when predicting chemostat growth (**Figure 16B**): from 10,000 simulations with randomized k_{cat} values, over 99.9% showed the metabolic shift towards fermentation at a specific growth rate under 0.1 h^{-1} or no fermentation at all (**Figure S8** in **Paper IV**). These findings indicate that it is important to perform manual curation to the kinetic data (at least in the energy-generating pathways), as incorrect values from databases could introduce bias in our simulations.

4.2. The challenge of using absolute proteomics data

Now that I have introduced a proteomics dataset under multiple levels of stress (**Paper II**) and a method for integrating proteomics data (**Paper IV**), the final challenge is to apply the method on the dataset in order to study enzyme usage across experimental conditions. However, before proceeding it is relevant to address the variability of proteomics data, and the challenges that this imposes on our analysis, as on a first glance the proteomics data clusters depending on the measuring instrument rather than the phenotypic response of yeast (**Figure 9F**). In this section I go through **Paper V**, where a dataset of yeast grown at reference conditions in chemostat was generated to study the variability of the iBAQ technique for estimating protein abundances. Furthermore, alternative approaches for converting the MS intensities into protein abundances are assessed to find which one increases accuracy and precision of predictions.

4.2.1. Variability in proteomics: Golden slumbers

To study variability of proteomics data computed using the iBAQ approach, a proteomics dataset was generated with both biological replicates, i.e. samples from different cultivations grown under the same conditions and analyzed by the MS instrument at the same time; and different MS batches, i.e. samples from the same cultivation analyzed at different times in the MS instrument. *S. cerevisiae*, strain CEN.PK113-7D, was grown in triplicate in aerobic glucose-limited chemostats in minimal media at a dilution rate of 0.1 h^{-1}, and each cultivation was later analyzed on three separate runs of the MS instrument, with a time difference of 12 and 30 days. The previously introduced iBAQ approach [80] was used to estimate absolute abundances of an internal standard from known abundances of an external standard, and the SILAC approach [81] was used to infer the absolute abundances for all samples from the internal standard intensities. All experimental settings were kept the same as described in **Paper II**.

As all measurements in this dataset come from the same experimental conditions, for each protein the values across measurements should be the same. I tested whether this was true by separately comparing the biological replicates (**Figure 17A**) and the MS batches (**Figure**

17B). Overall, there is a good agreement across biological replicates, similar to what has been observed in previous studies [126]; however, batches analyzed at different time points exhibit a much larger variability between them, a phenomenon referred to as the "batch effect" that is observed in many types of omics data [127]. This is confirmed when performing a PCA on the data (**Figure 17C**), where samples cluster based on the MS batches. This is undesired, as ideally the larger source of variability should come from the biology and not from the measuring instrument.

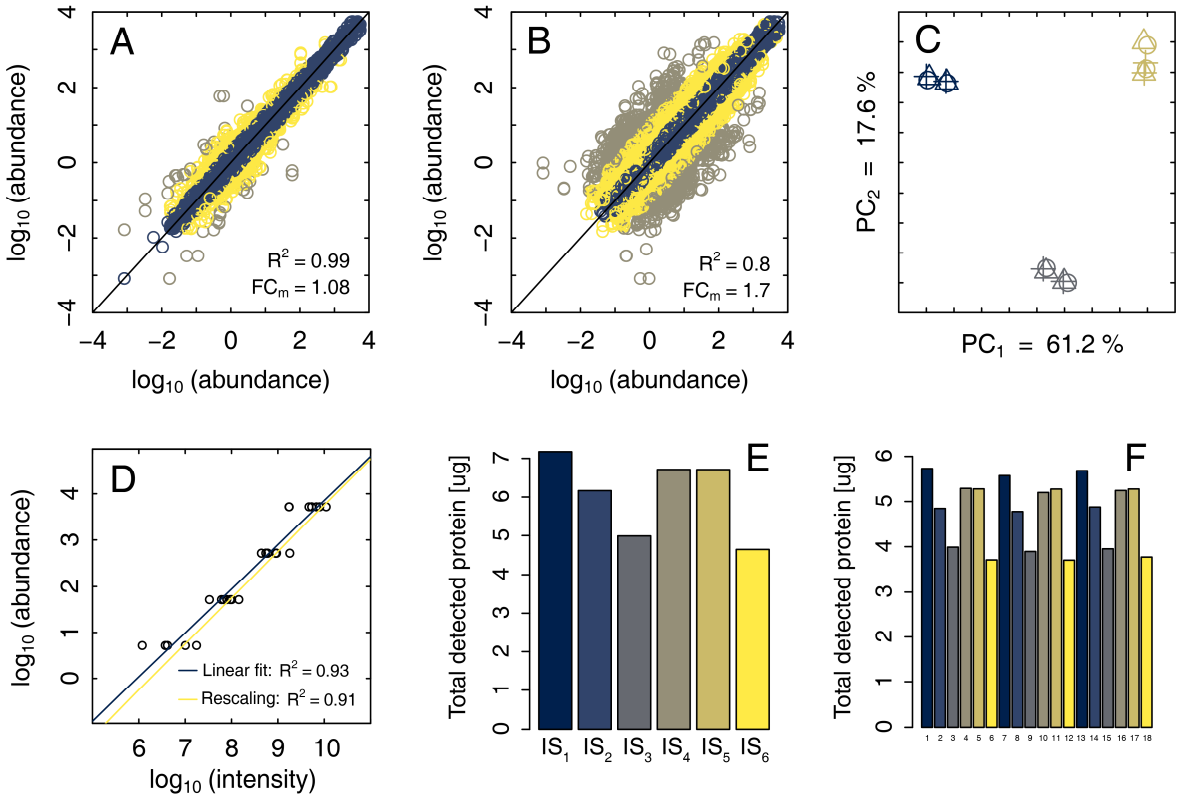

Figure 17: Variability of the proteomics data when using the iBAQ technique. (A-B) Variability between biological replicates (A) and batches (B). Blue circles correspond to fold changes between replicates lower than 2-fold, yellow between 2-fold and 10-fold, and grey above 10-fold. The coefficient of determination (R^2) and the median absolute fold change (FC_m) are shown. (C) PCA of all samples, indicating the amount of variability explained by each of the components. Different batches are shown with different colors, and different biological replicates with different shapes. (D) Known abundances of the external standard versus detected MS intensities. A linear fit to the data and a linear model based on scaling the data are shown. (E) Total detected protein for each of the 6 internal standards. (F) Total detected protein for each of the 18 samples. The colors match the corresponding internal standard used to compute the abundances.

A big influence on the batch variability might be the use of an external standard, as the detection of the proteins in the external standard carries a large degree of variability. Even though the UPS2 mix has only six levels of molar abundance, the detection by the MS instrument spans up to an order of magnitude of variability for each of the four detected levels (**Figure 17D**), in accordance to previous studies that use the iBAQ approach [80,128]. Note that as this data is used for building the standard curve in the log space, many curves can fit the data almost as well (e.g., both curves shown in **Figure 17D**). However, even a small change of slope can have a big influence on the final absolute values; in fact, using the optimal linear fit for each of the 6 external standard measurements leads to quite different

results for both internal standards (**Figure 17E**) and samples (**Figure 17F**) when we add up all predicted abundances. As in all cases the same amount of protein was injected into the MS instrument, we should expect the same amount to be detected. Therefore, I proceeded to test if the variability can be reduced by trying different approaches for rescaling the data so that it adds up to the same injected protein mass.

4.2.2. Increasing the quality of proteomics data: Little by little

In total I implemented three rescaling methods and compared them to the traditional iBAQ approach, referred to in the following as method 1. Method 2 rescaled the abundance values computed from method 1 so that the sum of all protein abundances would add up to the total injected protein mass. Method 3 is known as the total protein approach (TPA) [129] and is similar to method 2, but instead of using the abundances from method 1, it directly uses the detected MS intensities. Finally, method 4 is a variation of the TPA approach that first normalizes the MS intensities by the corresponding number of theoretical peptides that each protein has [130]. Note that methods 3 and 4 bypass the need of an external standard entirely. The full formalism for each method is available in **Paper V**.

In order to compare the performance of these methods, two important metrics that should be introduced here are accuracy and precision, as they are both routinely used in proteomics [131]. Accuracy refers to how far off the true value the prediction lies, i.e. if a prediction is consistently different than the expected value we can say the prediction is not accurate (or biased). Precision on the other hand refers to how variable said prediction is when the same measurement is repeated, i.e. if a measurement tends to take many different values, even if it is on average close to the expected value, we can say the prediction is imprecise (or non-reproducible). Note that a measurement can be inaccurate, imprecise, or both (**Figure 18**). In the following, I will look into accuracy and precision of each of the 4 mentioned methods.

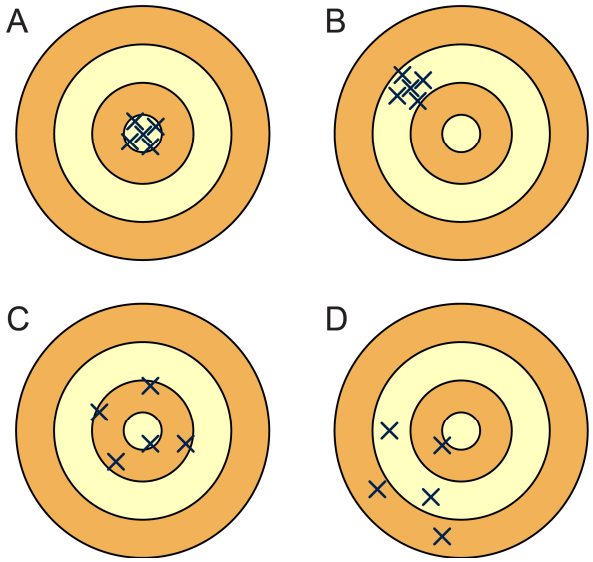

Figure 18: Exemplifying the concepts of accuracy and precision. (A) A measurement that is both accurate and precise. (B) A measurement that although precise is inaccurate. (C) A measurement that is on average accurate but imprecise. (D) A measurement that is neither accurate nor precise.

Accuracy was studied with two main metrics: UPS2 prediction error and ribosomal stoichiometry prediction error. For the first metric, the predicted abundances for the detected proteins in the external standard were compared to the known values in the UPS2 mix. This yielded similar performance for methods 1, 2 and 4, whereas method 3 performed significantly worse (**Figure 19A**). For the second metric, I looked into the stoichiometry of ribosomal proteins, which are known to be expressed in a 1-1 molar ratio [132]; therefore, comparing all molar estimates to the median value can be a good metric of accuracy [133]. It can again be observed that methods 1, 2 and 4 perform similarly, while method 3 has a lower performance (**Figure 19B**).

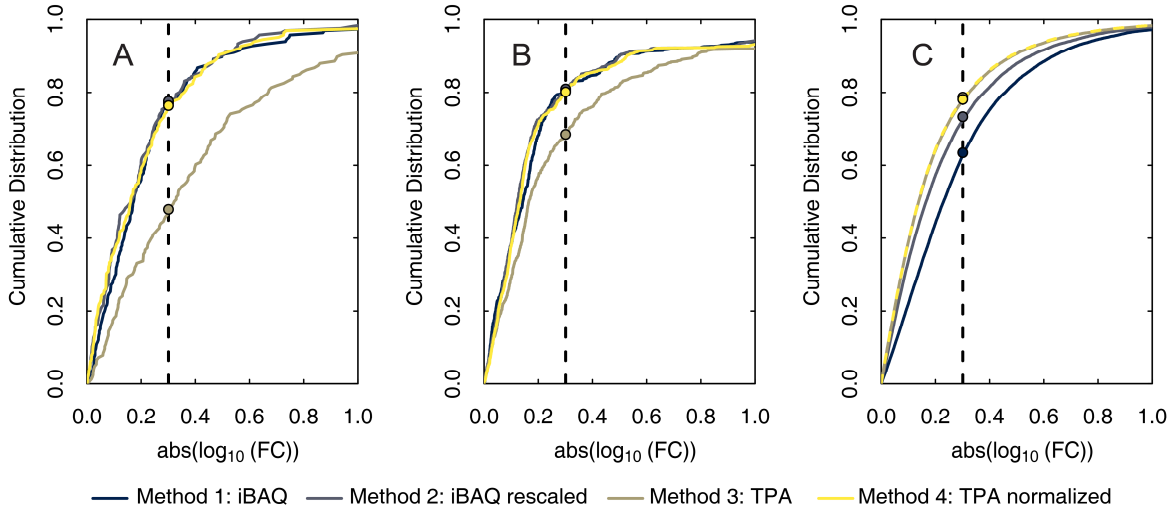

Figure 19: Accuracy and precision in a proteomics dataset after varying the scaling method. A fold change of 2 is indicated with a vertical segmented line. (A) Accuracy of predicting the external standard abundances (167 measurements). (B) Accuracy of predicting the ribosomal protein abundances (731 measurements). (C) Precision of predicting the data across batches (21,320 measurements).

Finally, precision was studied by performing the same comparison as in **Figure 17B** for all 4 methods. This comparison yielded that methods 3 and 4 perform better than methods 1 and 2 (**Figure 19C**). In particular, the original iBAQ method has around 40% of its protein estimates with a variability over a 2-fold within batches, whereas this drops to 20% in the case of the two methods that skip the external standard (methods 3 and 4). Considering all of the above, method 4 stands out as the best performing method for estimating protein abundances, as it is more accurate than method 3 and more precise than methods 1 and 2. Note that even though the external standard data was not used in this method, the linear models that convert the intensity data into absolute abundances are quite similar (**Figure 17D**).

It is important to mention that even though method 4 proved to be the best performing method, it is by no means a perfect method, with batch variability still being considerably larger than biological variability (**Figure 2** in **Paper V**). This clearly shows the limitations of working with absolute proteomics data, and that this variability should be accounted for when applying methods (e.g. GECKO) that use this type of data. In the next section I will proceed with an analysis on proteomics data and assess the effect of filtering out the fraction of the data that exhibits high variability. The data, together with the computational analysis in this section, are available at https://github.com/SysBioChalmers/reproduce.

4.3. Enzyme usage of *S. cerevisiae* during stress

Being aware of the inherent variability of the absolute proteomics data used, we proceed to the final part of this chapter, where I go through **Paper VI**. Here, enzyme-constrained models with proteomics data were generated for each of the experimental conditions from **Paper II**, filtering out the data that was excessively variable. The models had improved predicting performance compared to alternative approaches and allowed studying enzyme usage both at the global and stress-specific levels.

4.3.1. Condition-specific models of yeast at increasing levels of stress

For each experimental condition from the stress dataset, a condition-specific model was created using GECKO and the measured proteomics data. The data was first filtered to avoid the introduction of a bias in our analysis, leaving out any protein that had an average fold change over 2 between the reference conditions of the two datasets. The reasoning behind this is that if a value fluctuates in several folds between measurements of the same biological condition, it should not be trusted. This left 1746 proteins from the original 2318 measured proteins. Additionally, measurements from the subunits in the oxidative phosphorylation complexes were rescaled to be proportional to their median values.

For all enzymes in the model that had a direct match in the filtered proteomics data, an upper bound was set based on the average value and standard deviation across triplicates. Depending on the condition, this was possible for between 254 and 340 enzymes. Note that for each condition, around 7 measurements were increased in an iterative fashion to allow functional models, and the corresponding proteins were hence not accounted for in any further analysis. Finally, for all unmeasured enzymes in the model, the shared pool assumption was used (**Figure 14B**) assuming an average saturation factor of 50%. Considering both individual enzyme constraints and the shared enzyme pool, the condition-specific models accounted on average for 0.33 g/gDW of protein, i.e. 57.5% of the total protein content.

The created condition-specific models were simulated using pFBA and minimizing the carbon uptake [g/gDWh] subject to a growth of 0.1 h^{-1} (the dilution rate of all chemostats). Also, for each condition the NGAM value was increased so that the exchange rates of glucose, oxygen, CO_2 and ethanol would fit the measured values, as it has been shown that the stress response of yeast is tightly associated to an increase in energy maintenance [82]. These simulations were compared to corresponding simulations of the original metabolic model and an enzyme-constrained model built with no proteomics data (using the shared pool formalism presented in **Section 4.1.2**). Overall, the condition-specific models showed the best performance amongst all three approaches (**Figure 20A**).

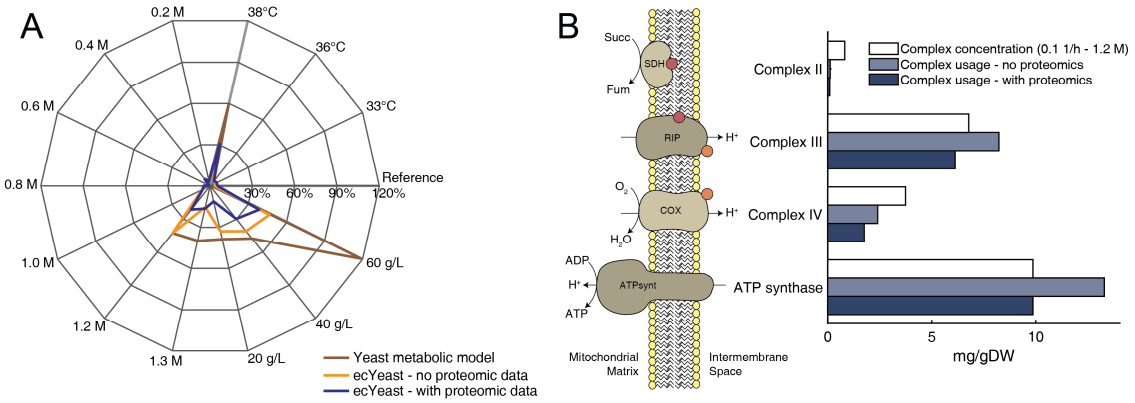

Figure 20: Condition-specific models. (A) Average error when predicting physiological data (glucose, O_2, CO_2 and ethanol exchange rates) with the original GEM of yeast, *ecYeast* without proteomics data, and *ecYeast* with proteomics data. (B) Enzyme usage of the complexes in oxidative phosphorylation at osmotic stress conditions of 1.2 M of NaCl. A diagram of oxidative phosphorylation is shown for reference.

As already covered in **Section 4.1.3**, a purely metabolic model is not able to choose fermentation as an energy-generating pathway, therefore the metabolic model fails to predict ethanol at high temperature (**Figure 9A**) or osmotic levels (**Figure 9B**). Due to the same reason, the metabolic model predicts only consumption of ethanol for conditions of ethanol stress, where there should be a mixed consumption of glucose and ethanol (**Figure 9C**). This is because consuming ethanol yields the highest ATP/carbon yield, even though it relies solely on respiration. In turn, the condition-specific enzyme-constrained models do capture all these mentioned phenomena.

Comparing to the enzyme-constrained models with no proteomics data, overall predictions are also better when incorporating condition-specific proteomics data (**Figure 20A**). This is because some of the metabolic responses of yeast to stress are not due to a limitation in the total protein content, but due to limitations of some specific protein abundances. For instance, at high levels of osmotic stress, oxidative phosphorylation complexes are not highly abundant, causing a saturation of respiration that forces fermentation (**Figure 20B**) without increasing the glucose uptake, something that would not be possible to predict with the molecular crowding formalism.

4.3.2. Enzyme usage response of yeast under stress

With the condition-specific models it is now possible to study enzyme usage at the genome-scale, both in absolute terms or as percentages of the measured values. It can be seen that some enzymes have a usage close to 100%; however, most enzymes tend to have usage values far below the measured abundances (**Figure 21A**). As this could be an effect of the k_{cat} values assigned (given that GECKO prefers data from more efficient enzymes), I examined any enzyme that showed a positive correlation of their enzyme usage with their own abundance ($R^2 > 0.9$) across all experimental conditions, pointing to a constant percentual enzyme usage. In total, 35 enzymes were in this group (**Table S2** in **Paper VI**), with an overrepresentation of oxidative phosphorylation proteins. The fact that these 35 enzymes show a constant percentual enzyme usage implies that the cell tightly regulates the expression of those enzymes, i.e. they are important for cell growth and could be either

translationally or transcriptionally controlled. The latter has been experimentally shown for some of these enzymes [134] but remains to be confirmed for most of them.

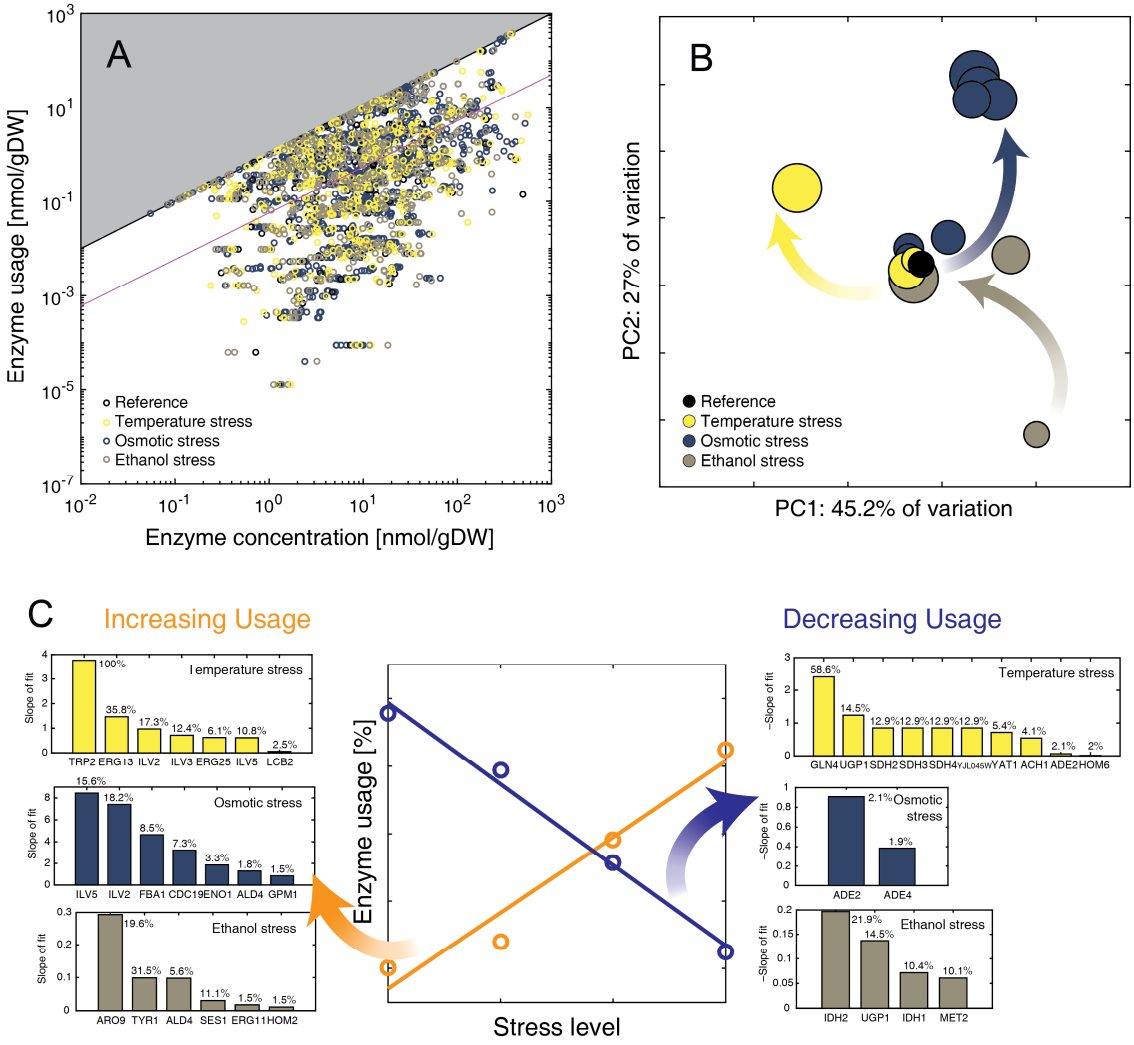

Figure 21: Enzyme usage trends. (A) Enzyme usage versus concentration for all enzymes in yeast, color coded by the type of stress. The obscured area refers to a region of infeasibility, as usage must remain below the abundance. A linear fit to all data is shown with a straight line. (B) PCA of percentual enzyme usages. Colors represent the stress types and marker sizes represent the stress level. The amount of variability that each component represents is indicated. (C) Enzymes showing either an increase (left side) or decrease (right side) in usage as the level of a specific type of stress increases, ranked by the slope of the fitted linear model. The maximum enzyme usage observed is shown on top of each bar.

Enzyme usage can also be studied among stress levels. It is less biased by the "batch effect" previously observed (**Figure 21B**), as it considers both the protein abundances and the metabolic flux. Additionally, it is a useful layer of information that connects proteomics with fluxomics, capturing metabolic responses that proteomics do not, while staying within the enzyme level. This is exemplified by the fact that it displays a clear convergent trend towards the reference condition for the ethanol stress chemostats (**Figure 21B**), a behavior that is not observed in the proteomics data (**Figure 9F**), but it is observed in the flux predictions, as an increased concentration of ethanol in the media pushes yeast towards increased glucose consumption (**Figure 9C**).

Finally, enzyme usage can also be studied within stress types. For this, I looked for enzymes that show a positive or negative correlation of their percentual usage with the stress level (R^2 > 0.9). These enzymes are expected to be important for stress tolerance: for instance, if an enzyme shows a positive correlation, it means that either the metabolic flux is systematically increasing, and/or the enzyme abundance is systematically decreasing. Either way, it points to an enzyme that plays a more important role as the stress increases, and could be a good target for studies to improve tolerance to said stress. The analysis revealed several enzymes that increased or decreased consistently with the stress levels (**Figure 21C**). From them, some were consistent with previous experimental studies that showed that engineering the same pathway could increase stress tolerance, such as ergosterol biosynthesis during heat stress [135]. However, most of them remain to be tested to assess if modified expression of the corresponding genes could yield a more tolerant strain to the corresponding stress.

In conclusion, using GECKO to constrain GEM simulations with proteomics data allows to study enzyme usage at the genome-scale, to find previously unknown trends in enzyme usage, both at the global level and at increasing levels of stress. The computational analysis in this section is available at https://github.com/SysBioChalmers/ecModels.

5. Conclusion

In this thesis, I have explored different approaches for combining genome-scale modeling with experimental omics data. I started by reviewing the metabolic modeling field and how it has been applied in *S. cerevisiae* (**Paper I**), to show that two challenges in the field are to properly evaluate the quality of GEMs and consistent omics data integration. I then proceeded to establish foundation stones for high quality GEMs and omics data. For the former, I introduced a version control strategy that records development of the model in a sustainable way (**Section 2.1**) and implemented it for the consensus GEM of yeast. For the latter, I introduced a multi-omics dataset of yeast grown under several conditions of stress (**Paper II**), showing that yeast re-arranges its metabolic distribution, biomass composition and protein levels in a drastic way as the level of stress increases.

Most of the thesis dealt with integration of lipid data and enzyme data into GEMs. To integrate lipid data, I have introduced SLIMEr (**Paper III**). By implementing SLIMEr on yeast, I showed that I could accurately represent amounts of lipid species, analyze the flexibility of the resulting distribution, and compute energy costs of moving from one metabolic state to another. To integrate enzyme data in GEMs, I have introduced GECKO (**Paper IV**). By implementing GECKO on yeast, I showed that the new model could correctly describe yeast physiology at high growth rates, which are conditions that entail high enzymatic demand, and compute usage among enzymes and metabolic pathways. GECKO also allows to directly integrate quantitative proteomics data; by doing so flux variability of the model was significantly reduced.

I also assessed the quality of the protein quantification technique employed throughout this thesis (**Paper V**). I presented current limitations of this technique, such as batch variability being higher than biological variability. Moreover, I introduced a simple normalization and rescaling approach that performs as accurately yet more precisely than methods that rely on external standards. Finally, I used GECKO to integrate the proteomics data from the stress dataset (**Paper VI**). I showed that the generated condition-specific models could better predict the metabolic stress response compared to previous modeling approaches. Furthermore, the model gave insight into the genome-wide distribution of usage both at the global and stress-specific levels, finding enzymes that play key-roles in different pathways inside the cell at conditions of high energy demand.

6. Future perspectives

While systems biology has become a well-established field, it also remains emergent in many ways: new experimental methods yield new types of omics data and insight; as computational biologists we then need to continuously update the way we analyze data and interpret simulation results. In this section I provide some views on where I think the field is moving to, with respect to the themes covered by this thesis.

6.1. Reproducible software in biology: Modeling with advantage

Reproducibility is a foundation stone from which all research should be built upon [62]. In the case of computational modeling, version control tools like Git and hosting services like GitHub are essential for achieving both traceability and reproducibility. Computational biology has been slow to adopt these good software development practices, but since a few years this has been changing, with more groups realizing the value of version control for reproducible research [127,136]. Version control is not only helpful when collaborating, but also when working alone, as we are always collaborating with our future self. On several occasions throughout my PhD I had to re-run analysis that I had done a few months before, and my experience was always better when the code was version-controlled, as I would be confident that no undesired changes had been introduced.

All the computational tools presented in this thesis are available in GitHub, so that most of the performed analysis can be replicated by any user. This sustainable approach to code development does however come with some minor challenges: it can become hard to maintain, as software packages and toolboxes are often being updated, which means that we need to adapt our code appropriately to maintain its proper functioning, or use additional tools such as dependency tracking and/or container platforms. Despite these hurdles, I envision that with proper investment and support, version control will become not just a standard, but a requirement for anyone performing computational analysis. Particularly, I also hope the practice of keeping track of GEMs with version control will spread to become a required feature by the community.

Moving forward in the sustainable development of GEMs, I believe that the idea of standardized testing of models needs to be implemented at a larger scale. In this regard, a recently developed tool for metabolic model testing (**Paper VIII**) is starting to be used by GEMs as an automatic tool for testing model quality, fully compatible with the version control strategy outlined in **Section 2.1**, acting as an additional test when changes are requested to be integrated in the model. I then hope this tool becomes the main metric to measure GEM quality, and a tool to lead development teams that curate these models.

6.2. Lipid constraints: Growing strong

The addition of SLIME reactions to a GEM opens many new paths for testing the flexibility of lipid metabolism in a quantitative way. As lipid quantification becomes cheaper and easier to parallelize, a natural extension of SLIMEr would be to further constrain the lipid distribution, this time with the acyl chain distribution for each lipid class. This would be especially relevant for metabolic engineering projects focused on reorganizing the acyl chain configuration within specific lipid classes [137]. It might as well be of value to include more lipid species in *yeast-GEM*, considering that only 80% mass-wise of the lipid content is currently covered, and only C16, C18, C24 and C26 acyl chains are present in the model.

It is also of interest to study the benefit of this approach for other organisms, especially in other yeast species that display a much higher lipid content, such as *Yarrowia lipolytica* [6]. In fact, SLIMEr has recently been implemented in *Rhodotorula toruloides* [138]. A challenge to implement the approach in even more organisms is the non-standardized way that lipid names are often annotated in GEMs, as this requires intensive manual work for properly creating the generic pseudo-metabolites (backbones and acyl chains) and adding the SLIME reactions. Efforts in including standardized IDs in GEMs [34] are fundamental for making this step as seamless as possible.

6.3. Enzyme constraints: From soft to hard and back

Enhancing a GEM with enzyme constraints has shown improved prediction capabilities, and I think this approach can be developed even further. So far, I have presented two variations of the approach: a first one in which no proteomics data is used and only a single constraint is applied, and a second one in which proteomics data is used as constraints in single enzymes. I have defined these as "soft" and "hard" approaches, respectively (**Section 1.3.4**). Note that the latter approach requires a couple of flexibilization steps in the proteomics data (**Paper VI**), which decreases the coverage of the method, as the flexibilized enzymes are removed from posterior analysis. This method could then be adapted to avoid excessive flexibilization, by for instance creating constraints for subgroups of enzymes, e.g. pathways, instead of single enzymes. Alternatively, the soft approach could be further used by comparing enzyme usage predictions to the measured abundances.

Finally, note that in both soft and hard approaches there is a constraint on the shared metabolite pool of all unmeasured enzymes (**Section 4.1.2**), which can affect simulations considerably (**Section 4.1.3**). I envision an even "softer" approach, in which no actual constraint is imposed (not even on the shared metabolite pool), and instead the model is run to compute enzyme usages, to compare them later to proteomics data. All these approaches could offer new insight and I believe should be assessed in a future study.

6.4. The importance of good data: House of cards

Thanks to its ease of use, GECKO has already been used in a genome-scale model of *B. subtilis* [139] and even more models are currently being developed for bacteria, other yeasts and even human cell lines (**Paper XII**). However, something that has become clear after the analysis in this thesis is that the developed approach depends to a large extent on the different types of data used: the kinetic data obtained from literature, the measured proteomics data,

the biochemical knowledge used to build the original metabolic network, and the biomass composition used in simulations. In the following I address the limitations that this imposes for moving forward with GECKO and how these limitations could be addressed in future studies.

Regarding kinetic data, I have previously mentioned that simulations can majorly be affected by the choice of k_{cat} values, and that even for the model organism *S. cerevisiae* the available kinetic data in the literature is scarce (**Figure 15**), which means that for other less studied organisms the scarcity will be even higher. Hence, studies that could generate kinetic information at the genome-scale in a systematic way would be of the upmost usefulness to the modeling community [140]. This however presents a set of challenges of its own, as the experimental setup for measuring kinetic parameters in a high-throughput way is non-trivial, and values vary among organisms and experimental conditions. Furthermore, enzymes inside the cell are subject to evolutionary pressure, which over time could affect some of their kinetic parameters, rendering the original measurements outdated. The latter however might be more prevalent in pathways that the cell does not routinely use (e.g. resistance to toxic compounds) and might not be a huge deterrent in our analysis.

Regarding proteomics, in this thesis I have outlined the challenges this type of omics data presents when it comes to reproducibility across measurements (**Paper V**). These challenges have also become evident in a recent study [141] that integrated different datasets of MS-generated proteomics measurements, which yielded large variability across studies. To cope with these challenges, in this thesis I explored different scaling methods (**Paper V**) and filtered out excessively variable data (**Paper VI**) before applying any quantitative analysis. However, it could very well be the case that unreliable data is still included in our analysis. New methods for measuring absolute proteomics abundance that exhibit improved precision [142] should become the *de-facto* way of generating data, and additional control quality checks should be put in place to guarantee the robustness of the data used.

Regarding biochemical knowledge, it is important to consider that GEMs are built based on our current knowledge of the roles of different genes, and the activities of different enzymes. However, this knowledge remains incomplete, with a big fraction of experimentally detected metabolites not present in GEMs [143]. Additionally, due to this missing knowledge a big part of metabolic networks often contains blocked reactions. To address the latter, gap-filling techniques have been developed that can reduce the problem to some extent [144]. Nonetheless, further efforts towards a fully characterized network are still needed.

A final way of improving predictions could be with a better-defined biomass composition. Using GECKO to infer enzyme usages predicted that many of the detected enzymes had zero usage for all conditions (**Paper VI**), which is unlikely to be a realistic biological behavior, i.e. the cell producing a big part of its metabolic proteome for no apparent reason. What is more likely is that we have limited knowledge of the metabolic network, especially when it comes to the biomass requirements: As many pathways in the metabolic network are for the production of specific compounds that are currently not defined as part of the biomass pseudo-reaction, it is very likely that if we would re-formulate the biomass pseudo-reaction to account for more components, the pathways generating said components would become unblocked and hence enzymes in those pathways would have non-zero usage. A better characterization of the biomass composition would then yield more meaningful simulations [57].

6.5. Systems biology: Over the hills and far away

What is the ultimate goal of systems biology? A straightforward (but perhaps naïve) answer is "to simulate the complete behavior of a cell". However, I believe it is relevant to consider the aphorism "*all models are wrong, but some are useful*" [145]. A full depiction of the cell would be extremely hard to attain, and perhaps not significantly more useful than a simpler representation. Introduced in 2012, whole-cell modeling, i.e. modeling all processes in the cell with a single model [109], seemed promising at the time; however, numerous limitations in terms of parameter estimation have prevented the field to advance in that direction [146]. Many of these parameters might not be possible to measure, or might even be changing due to evolutionary pressure, making the goal of a fully encompassing model an elusive objective.

I believe a simpler approach in which we slowly build our understanding of the cell by integrating one layer of data at the time is a better way to move forward. Particularly, in this thesis I have introduced GECKO, which allows connecting two important levels of information in the cell (metabolism and enzymes), to then infer enzyme usage. Here, it is vital to note that many additional processes affect the metabolic rates in combination with the enzyme levels, such as the degree of saturation of the enzyme due to the substrate levels, the thermodynamics of the reaction, inhibitory/activation effects, and the influence of the environment [147]. In fact, my simulations show that enzyme usage seems to remain rather low for many enzymes in metabolism (**Figure 21A**), which has been observed experimentally for bacteria [148]. Therefore, this usage should be interpreted as a *minimum* enzyme usage, or what has also been referred to as "minimal enzyme demand" [149].

The next step of this approach would then be to account for more types of constraints, as evolution pushes biological systems to be multi-constrained by different intracellular abundances and processes. There are a plethora of studies assessing the benefits of different types of constraints on GEMs, such as substrate/product levels [149], gene expression rates [107], metabolite transport/diffusion [150] and Gibbs energy dissipation [122]. An integrative study that would assess the benefits of combining one or more of these approaches with GECKO would be in my opinion a logical next step. In particular, given that not all of the experimental data in this study was used for integration with GEMs (**Paper II**), accounting for transcriptomics data and protein degradation seems to be an attractive way of moving forward.

And then what? As introduced in **Chapter 1**, metabolic engineering has recently benefitted from efficient techniques for fine-tuning the levels of gene expression [10]. I envision the modeling framework presented in this thesis as a tool for predicting the optimal levels of said expression: The GECKO approach already yields enzyme usage predictions, and by accounting for additional processes such as the mRNA/enzyme relationship [107] and the enzyme concentration/usage relationship [149], we would have a model that predicts the full effect of gene expression levels on metabolic fluxes. This model could guide experimentalists to decide which genes to target and what levels of over-expression / down-regulation to use, for them to achieve the desired phenotypic traits of their mutant strains. Ultimately, I hope this thesis inspires an increased use of model-based design in metabolic engineering.

7. References

1. Mantai L, Dowling R. Supporting the PhD journey: insights from acknowledgements. International Journal for Researcher Development. Emerald Group Publishing Limited; 2015;6: 106–121. doi:10.1108/ijrd-03-2015-0007
2. Festel G. Industrial biotechnology: Market size, company types, business models, and growth strategies. Industrial Biotechnology. Mary Ann Liebert, Inc.; 2010;6: 88–94. doi:10.1089/ind.2010.0006
3. McGovern PE, Zhang J, Tang J, Zhang Z, Hall GR, Moreau RA, et al. Fermented beverages of pre- and proto-historic China. Proceedings of the National Academy of Sciences. National Academy of Sciences; 2004;101: 17593–17598. doi:10.1073/pnas.0407921102
4. Tang WL, Zhao H. Industrial biotechnology: Tools and applications. Biotechnology Journal. John Wiley & Sons, Ltd; 2009. pp. 1725–1739. doi:10.1002/biot.200900127
5. Hong KK, Nielsen J. Metabolic engineering of Saccharomyces cerevisiae: A key cell factory platform for future biorefineries. Cellular and Molecular Life Sciences. SP Birkhäuser Verlag Basel; 2012. pp. 2671–2690. doi:10.1007/s00018-012-0945-1
6. Beopoulos A, Cescut J, Haddouche R, Uribelarrea JL, Molina-Jouve C, Nicaud JM. Yarrowia lipolytica as a model for bio-oil production. Progress in Lipid Research. Pergamon; 2009. pp. 375–387. doi:10.1016/j.plipres.2009.08.005
7. Botstein D, Chervitz SA, Cherry JM. Yeast as a Model Organism. Science. 1997;277: 1259–1260. doi:10.1126/science.277.5330.1259
8. Nielsen J. It Is All about Metabolic Fluxes. Journal of Bacteriology. American Society for Microbiology Journals; 2003. pp. 7031–7035. doi:10.1128/JB.185.24.7031-7035.2003
9. Stephanopoulos GN, Aristidou AA, Nielsen J. Metabolic engineering : principles and methodologies. San Diego: Academic Press. 1998. doi:10.1016/B978-0-12-666260-3.50019-4
10. Farzadfard F, Perli SD, Lu TK. Tunable and multifunctional eukaryotic transcription factors based on CRISPR/Cas. ACS Synthetic Biology. 2013;2: 604–613. doi:10.1021/sb400081r
11. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review. 1956;63: 81–97. doi:10.1037/h0043158
12. Kitano H. Systems biology: a brief overview. Science. American Association for the Advancement of Science; 2002;295: 1662–4. doi:10.1126/science.1069492
13. Palsson BØ. Systems Biology: Constraint-based Reconstruction and Analysis. Cambridge, UK: Cambridge University Press; 2015. doi:10.1017/CBO9781139854610
14. Stalidzans E, Seiman A, Peebo K, Komasilovs V, Pentjuss A. Model-based metabolism design: constraints for kinetic and stoichiometric models. Biochemical Society Transactions. 2018;46: 261–267. doi:10.1042/BST20170263
15. Bordbar A, Monk JM, King ZA, Palsson BØ. Constraint-based models predict metabolic and associated cellular functions. Nature reviews Genetics. 2014;15: 107–20. doi:10.1038/nrg3643
16. Kerkhoven EJ, Lahtvee PJ, Nielsen J. Applications of computational modeling in metabolic engineering of yeast. FEMS Yeast Research. 2014;15: 1–13. doi:10.1111/1567-1364.12199
17. Almquist J, Cvijovic M, Hatzimanikatis V, Nielsen J, Jirstrand M. Kinetic models in industrial biotechnology - Improving cell factory performance. Metabolic Engineering. Elsevier; 2014;24: 38–60. doi:10.1016/j.ymben.2014.03.007
18. Sánchez BJ, Pérez-Correa JR, Agosin E. Construction of robust dynamic genome-scale metabolic

model structures of Saccharomyces cerevisiae through iterative re-parameterization. Metabolic Engineering. 2014;25: 159–173. doi:10.1016/j.ymben.2014.07.004

19. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabási AL, et al. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proceedings of the National Academy of Sciences of the United States of America. 2007;104: 12663–12668. doi:10.1073/pnas.0609845104
20. Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proceedings of the National Academy of Sciences. National Academy of Sciences; 2000;97: 5528–5533. doi:10.1073/pnas.97.10.5528
21. Lewis NE, Nagarajan H, Palsson BØ. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nature Reviews Microbiology. Nature Publishing Group; 2012;10: 291–305. doi:10.1038/nrmicro2737
22. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U. Multidimensional optimality of microbial metabolism. Science. 2012;336: 601–604. doi:10.1126/science.1216882
23. Varma A, Palsson BØ. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Applied and Environmental Microbiology. 1994;60: 3724–3731.
24. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models: the COBRA Toolbox v3.0. ArXiV preprint. 2017; 1710.04038.
25. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. BMC Systems Biology. BioMed Central; 2013;7: 74. doi:10.1186/1752-0509-7-74
26. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nature Biotechnology. 2010;28: 245–248. doi:10.1038/nbt.1614
27. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Molecular systems biology. 2010;6: 390. doi:10.1038/msb.2010.47
28. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering. 2003;5: 264–276. doi:10.1016/j.ymben.2003.09.002
29. Schellenberger J, Palsson BØ. Use of randomized sampling for analysis of metabolic networks. The Journal of biological chemistry. American Society for Biochemistry and Molecular Biology; 2009;284: 5457–61. doi:10.1074/jbc.R800048200
30. Megchelenbrink W, Huynen M, Marchiori E. optGpSampler: An improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. PLoS ONE. Public Library of Science; 2014;9: e86587. doi:10.1371/journal.pone.0086587
31. Oberhardt MA, Palsson BØ, Papin JA. Applications of genome-scale metabolic reconstructions. Molecular systems biology. 2009;5: 320. doi:10.1038/msb.2009.77
32. Österlund T, Nookaew I, Bordel S, Nielsen J. Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. BMC systems biology. 2013;7: 36. doi:10.1186/1752-0509-7-36
33. Ghaffari P, Mardinoglu A, Asplund A, Shoaie S, Kampf C, Uhlen M, et al. cancer cell lines through genome-scale metabolic modeling. Scientific Reports. 2015;5. doi:10.1038/srep08183
34. Dräger A, Palsson BØ. Improving collaboration by standardization efforts in systems biology. Frontiers in Bioengineering and Biotechnology. 2014;2: 1–20. doi:10.3389/fbioe.2014.00061
35. Gevorgyan A, Poolman MG, Fell DA. Detection of stoichiometric inconsistencies in biomolecular models. Bioinformatics. Oxford University Press; 2008;24: 2245–2251. doi:10.1093/bioinformatics/btn425
36. Wagner A, Fell DA. The small world inside large metabolic networks. Proceedings Biological sciences / The Royal Society. 2001;268: 1803–1810. doi:10.1098/rspb.2001.1711
37. Sauer U. Metabolic networks in motion: 13C-based flux analysis. Molecular systems biology. 2006;2. doi:10.1038/msb4100109
38. Basler G. Computational prediction of essential metabolic genes using constraint-based approaches. Methods in Molecular Biology. 2015;1279: 183–204. doi:10.1007/978-1-4939-2398-4_12
39. Saha R, Chowdhury A, Maranas CD. Recent advances in the reconstruction of metabolic models and integration of omics data. Current Opinion in Biotechnology. Elsevier Ltd; 2014;29: 39–45.

doi:10.1016/j.copbio.2014.02.011

40. Mann M, Kulak NA, Nagaraj N, Cox J. The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. Molecular Cell. Elsevier; 2013. pp. 583–590. doi:10.1016/j.molcel.2013.01.029
41. Förster J, Famili I, Palsson BØ, Nielsen J. Genome-Scale Reconstruction of the Saccharomyces Cerevisie Metabolic Network. Genome Research. 2003;13: 244–253. doi:10.1101/gr.234503.complex
42. Nookaew I, Olivares-Hernández R, Bhumiratana S, Nielsen J. Genome-scale metabolic models of Saccharomyces cerevisiae. Methods in Molecular Biology. Totowa, NJ: Humana Press; 2011;759: 445–463. doi:10.1007/978-1-61779-173-4_25
43. Österlund T, Nookaew I, Nielsen J. Fifteen years of large scale metabolic modeling of yeast: developments and impacts. Biotechnology Advances. Elsevier Inc.; 2012;30: 979–988. doi:10.1016/j.biotechadv.2011.07.021
44. Herrgård MJ, Swainston N, Dobson PD, Dunn WB, Arga KY, Arvas M, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. Nature biotechnology. 2008;26: 1155–60. doi:10.1038/nbt1492
45. Dobson PD, Smallbone K, Jameson D, Simeonidis E, Lanthaler K, Pir P, et al. Further developments towards a genome-scale metabolic model of yeast. BMC systems biology. 2010;4: 145. doi:10.1186/1752-0509-4-145
46. Heavner BD, Smallbone K, Barker B, Mendes P, Walker LP. Yeast 5 - an Expanded Reconstruction of the Saccharomyces Cerevisiae Metabolic Network. BMC systems biology. 2012;6: 1. doi:10.1186/1752-0509-6-55
47. Heavner BD, Smallbone K, Price ND, Walker LP. Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. Database. 2013;2013: bat059. doi:10.1093/database/bat059
48. Aung HW, Henry SA, Walker LP. Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. Industrial biotechnology. 2013;9: 215–228. doi:10.1089/ind.2013.0013
49. Sánchez B, Li F, Lu H, Kerkhoven E, Nielsen J. SysBioChalmers/yeast-GEM: yeast 8.0.0. Zenodo. 2018. doi:10.5281/ZENODO.1495477
50. Duarte NC, Herrgård MJ, Palsson B. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Research. 2004;14: 1298–1309. doi:10.1101/gr.2250904
51. Kuepfer L, Sauer U, Blank LM. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. Genome research. 2005;15: 1421–30. doi:10.1101/gr.3992505
52. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, Cheevadhanarak S, et al. The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: A scaffold to query lipid metabolism. BMC Systems Biology. 2008;2: 71. doi:10.1186/1752-0509-2-71
53. Mo ML, Palsson BØ, Herrgård MJ. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. BMC systems biology. 2009;3: 37. doi:10.1186/1752-0509-3-37
54. Zomorrodi AR, Maranas CD. Improving the iMM904 S. cerevisiae metabolic model using essentiality and synthetic lethality data. BMC systems biology. BioMed Central Ltd; 2010;4: 178. doi:10.1186/1752-0509-4-178
55. Chowdhury R, Chowdhury A, Maranas CD. Using gene essentiality and synthetic lethality information to correct yeast and CHO cell genome-scale models. Metabolites. Multidisciplinary Digital Publishing Institute; 2015;5: 536–570. doi:10.3390/metabo5040536
56. Pereira R, Nielsen J, Rocha I. Improving the flux distributions simulated with genome-scale metabolic models of Saccharomyces cerevisiae. Metabolic Engineering Communications. Elsevier; 2016;3: 153–163. doi:10.1016/j.meteno.2016.05.002
57. Dikicioglu D, Kırdar B, Oliver SG. Biomass composition: the “elephant in the room” of metabolic modelling. Metabolomics. 2015;11: 1690–1701. doi:10.1007/s11306-015-0819-2
58. Dikicioglu D, Pir P, Onsan ZI, Ulgen KO, Kirdar B, Oliver SG. Integration of metabolic modeling and phenotypic data in evaluation and improvement of ethanol production using respiration-deficient mutants of Saccharomyces cerevisiae. Applied and Environmental Microbiology. 2008;74: 5809–5816. doi:10.1128/AEM.00009-08
59. Bro C, Regenberg B, Förster J, Nielsen J. In silico aided metabolic engineering of Saccharomyces cerevisiae for improved bioethanol production. Metabolic Engineering. 2006;8: 102–111.

doi:10.1016/j.ymben.2005.09.007

60. Hanly TJ, Henson MA. Dynamic model-based analysis of furfural and HMF detoxification by pure and mixed batch cultures of S. cerevisiae and S. stipitis. Biotechnology and Bioengineering. 2014;111: 272–284. doi:10.1002/bit.25101

61. O'Brien EJ, Monk JM, Palsson BØ. Using genome-scale models to predict biological capabilities. Cell. Elsevier Inc.; 2015;161: 971–987. doi:10.1016/j.cell.2015.05.019

62. Roger D Peng. Reproducible Research in Computational Science. Science. 2011;334: 1227. doi:10.1126/science.1213443

63. Chacon S, Straub B. Pro Git. 2nd ed. Apress; 2014. doi:10.1007/978-1-4842-0076-6

64. Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C. Automated genome annotation and metabolic model reconstruction in the SEED and model SEED. Methods in Molecular Biology. Humana Press, Totowa, NJ; 2013. pp. 17–45. doi:10.1007/978-1-62703-299-5_2

65. Dias O, Rocha M, Ferreira EC, Rocha I. Reconstructing genome-scale metabolic models with merlin. Nucleic Acids Research. Oxford University Press; 2015;43: 3899–3910. doi:10.1093/nar/gkv294

66. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic Acids Research. Oxford University Press; 2018;46: 7542–7553. doi:10.1093/nar/gky537

67. Karp PD, Latendresse M, Paley SM, Krummenacker M, Ong QD, Billington R, et al. Pathway tools version 19.0 update: Software for pathway/genome informatics and systems biology. Briefings in Bioinformatics. Oxford University Press; 2016;17: 877–890. doi:10.1093/bib/bbv079

68. Aite M, Chevallier M, Frioux C, Trottier C, Got J, Cortés MP, et al. Traceability, reproducibility and wiki-exploration for "à-la-carte" reconstructions of genome-scale metabolic models. PLoS Computational Biology. 2018;14: e1006146. doi:10.1371/journal.pcbi.1006146

69. Mack J. Semantic Commit Messages [Internet]. 2014. Available: https://seesparkbox.com/foundry/semantic_commit_messages

70. Driessen V. A successful Git branching model [Internet]. 2010. Available: https://nvie.com/posts/a-successful-git-branching-model/

71. Wisselink HW, Cipollina C, Oud B, Crimi B, Heijnen JJ, Pronk JT, et al. Metabolome, transcriptome and metabolic flux analysis of arabinose fermentation by engineered Saccharomyces cerevisiae. Metabolic Engineering. Elsevier; 2010;12: 537–551. doi:10.1016/j.ymben.2010.08.003

72. Attfield P V. Stress tolerance: The key to effective strains of industrial baker's yeast. Nature Biotechnology. Nature Publishing Group; 1997. pp. 1351–1357. doi:10.1038/nbt1297-1351

73. Çakir T, Patil KR, Onsan Z iIsen, Ulgen KO, Kirdar B, Nielsen J. Integration of metabolome data with metabolic networks reveals reporter reactions. Molecular systems biology. 2006;2: 50. doi:10.1038/msb4100085

74. Morano KA, Grant CM, Moye-Rowley WS. The response to heat shock and oxidative stress in saccharomyces cerevisiae. Genetics. Genetics; 2012;190: 1157–1195. doi:10.1534/genetics.111.128033

75. Alexandre H, Ansanay-Galeote V, Dequin S, Blondin B. Global gene expression during short-term ethanol stress in Saccharomyces cerevisiae. FEBS Letters. John Wiley & Sons, Ltd; 2001;498: 98–103. doi:10.1016/S0014-5793(01)02503-0

76. Blomberg A, Adler L. Physiology of Osmotolerance in Fungi. Advances in microbial physiology. Academic Press; 1992;33: 145–212. doi:10.1016/S0065-2911(08)60217-9

77. Villadsen J, Nielsen J, Lidén G. Bioreaction Engineering Principles. Springer; 2011. doi:10.1007/978-1-4419-9688-6

78. Chahrour O, Cobice D, Malone J. Stable isotope labelling methods in mass spectrometry-based quantitative proteomics. Journal of Pharmaceutical and Biomedical Analysis. Elsevier; 2015;113: 2–20. doi:10.1016/J.JPBA.2015.04.013

79. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology. Nature Publishing Group; 2008;26: 1367–1372. doi:10.1038/nbt.1511

80. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. Nature Research; 2011;473: 337–342. doi:10.1038/nature10098

81. Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M. Super-SILAC mix for quantitative

proteomics of human tumor tissue. Nature Methods. Nature Publishing Group; 2010;7: 383–385. doi:10.1038/nmeth.1446

82. Lahtvee P-J, Kumar R, Hallström B, Nielsen J. Adaptation to different types of stress converge on mitochondrial metabolism. Molecular Biology of the Cell. American Society for Cell Biology; 2016;27: 2505–2514. doi:10.1091/mbc.E16-03-0187

83. Nes WR, Nes WD. Lipids in Evolution. Boston, MA: Springer US; 1980. doi:10.1007/978-1-4684-3683-9

84. Feist AM, Palsson BO. The biomass objective function. Current Opinion in Microbiology. Elsevier Ltd; 2010;13: 344–349. doi:10.1016/j.mib.2010.03.003

85. Lachance J-C, Monk JM, Lloyd CJ, Seif Y, Palsson BO, Rodrigue S, et al. BOFdat: generating biomass objective function stoichiometric coefficients from experimental data. bioRxiv. Cold Spring Harbor Laboratory; 2018; 243881. doi:10.1101/243881

86. Asp NG. Dietary carbohydrates: Classification by chemistry and physiology. Food Chemistry. Elsevier; 1996. pp. 9–14. doi:10.1016/0308-8146(96)00055-6

87. Küenzi MT, Fiechter A. Changes in carbohydrate composition and trehalase-activity during the budding cycle of Saccharomyces cerevisiae. Archiv für Mikrobiologie. Springer-Verlag; 1969;64: 396–407. doi:10.1007/BF00417021

88. Ejsing CS, Sampaio JL, Surendranath V, Duchoslav E, Ekroos K, Klemm RW, et al. Global analysis of the yeast lipidome by quantitative shotgun mass spectrometry. Proceedings of the National Academy of Sciences. National Academy of Sciences; 2009;106: 2136–2141. doi:10.1073/pnas.0811700106

89. Mardinoglu A, Agren R, Kampf C, Asplund A, Uhlen M, Nielsen J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nature Communications. Nature Publishing Group; 2014;5: 3083. doi:10.1038/ncomms4083

90. Moreau RA. Lipid analysis via HPLC with a charged aerosol detector. Lipid Technology. Springer-Verlag; 2009;21: 191–194. doi:10.1002/lite.200900048

91. Khoomrung S, Chumnanpuen P, Jansa-Ard S, Ståhlman M, Nookaew I, Borén J, et al. Rapid Quantification of Yeast Lipid using Microwave-Assisted Total Lipid Extraction and HPLC-CAD. Analytical Chemistry. American Chemical Society; 2013;85: 4912–4919. doi:10.1021/ac3032405

92. Abdulkadir S, Tsuchiya M. One-step method for quantitative and qualitative analysis of fatty acids in marine animal samples. Journal of Experimental Marine Biology and Ecology. Elsevier; 2008;354: 1–8. doi:10.1016/J.JEMBE.2007.08.024

93. Khoomrung S, Chumnanpuen P, Jansa-Ard S, Nookaew I, Nielsen J. Fast and accurate preparation fatty acid methyl esters by microwave-assisted derivatization in the yeast Saccharomyces cerevisiae. Applied Microbiology and Biotechnology. Springer-Verlag; 2012;94: 1637–1646. doi:10.1007/s00253-012-4125-x

94. Kerkhoven EJ, Pomraning KR, Baker SE, Nielsen J. Regulation of amino-acid metabolism controls flux to lipid accumulation in Yarrowia lipolytica. npj Systems Biology and Applications. Nature Publishing Group; 2016;2: 16005. doi:10.1038/npjsba.2016.5

95. Henderson CM, Zeno WF, Lerno LA, Longo ML, Block DE. Fermentation Temperature Modulates Phosphatidylethanolamine and Phosphatidylinositol Levels in the Cell Membrane of Saccharomyces cerevisiae. Applied and Environmental Microbiology. 2013;79: 5345–5356. doi:10.1128/AEM.01144-13

96. Henderson CM, Lozada-Contreras M, Jiranek V, Longo ML, Block DE. Ethanol production and maximum cell growth are highly correlated with membrane lipid composition during fermentation as determined by lipidomic analysis of 22 saccharomyces cerevisiae strains. Applied and Environmental Microbiology. 2013;79: 91–104. doi:10.1128/AEM.02670-12

97. Brunk E, Sahoo S, Zielinski DC, Altunkaya A, Dräger A, Mih N, et al. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nature Biotechnology. Nature Publishing Group; 2018;36: 272–281. doi:10.1038/nbt.4072

98. Chan SHJ, Cai J, Wang L, Simons-Senftle MN, Maranas CD. Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. Bioinformatics. 2017;33: 3603–3609. doi:10.1093/bioinformatics/btx453

99. Sánchez B, Li F, Lu H, Kerkhoven E, Nielsen J. SysBioChalmers/yeast-GEM: yeast 8.1.0. Zenodo. 2018. doi:10.5281/ZENODO.1494212

100. Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T. Prediction of microbial growth rate versus

biomass yield by a metabolic network with kinetic parameters. PLoS Computational Biology. 2012;8: e1002575. doi:10.1371/journal.pcbi.1002575

101. Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect. PLoS Computational Biology. 2011;7: 1–8. doi:10.1371/journal.pcbi.1002018

102. Vazquez A, Beg QK, Demenezes MA, Ernst J, Bar-Joseph Z, Barabási A-L, et al. Impact of the solvent capacity constraint on E. coli metabolism. BMC systems biology. 2008;2: 7. doi:10.1186/1752-0509-2-7

103. Van Hoek MJ, Merks RM. Redox balance is key to explaining full vs. partial switching to low-yield metabolism. BMC Systems Biology. BioMed Central Ltd; 2012;6: 22. doi:10.1186/1752-0509-6-22

104. Nilsson A, Nielsen J. Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase. Scientific Reports. Nature Publishing Group; 2016;6: 22264. doi:10.1038/srep22264

105. Vazquez A, Oltvai ZN. Molecular crowding defines a common origin for the warburg effect in proliferating cells and the lactate threshold in muscle physiology. PLoS ONE. Public Library of Science; 2011;6: e19538. doi:10.1371/journal.pone.0019538

106. Thiele I, Jamshidi N, Fleming RMT, Palsson BO. Genome-scale reconstruction of escherichia coli's transcriptional and translational machinery: A knowledge base, its mathematical formulation, and its functional characterization. PLoS Computational Biology. 2009;5. doi:10.1371/journal.pcbi.1000312

107. O'Brien EJ, Palsson BØ. Computing the functional proteome : recent progress and future prospects for genome-scale models. Current Opinion in Biotechnology. Elsevier Ltd; 2015;34: 125–134. doi:10.1016/j.copbio.2014.12.017

108. Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, et al. Quantitative prediction of genome-wide resource allocation in bacteria. Metabolic Engineering. 2015;32: 232–243. doi:10.1016/j.ymben.2015.10.003

109. Karr JR, Sanghvi JC, MacKlin DN, Gutschow M V., Jacobs JM, Bolival B, et al. A whole-cell computational model predicts phenotype from genotype. Cell. 2012;150: 389–401. doi:10.1016/j.cell.2012.05.044

110. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for Penicillium chrysogenum. PLoS computational biology. 2013;9: e1002980. doi:10.1371/journal.pcbi.1002980

111. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015;15: 3163–3168. doi:10.1002/pmic.201400441

112. Machado D, Herrgård MJ, Rocha I. Stoichiometric Representation of Gene–Protein–Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction. PLOS Computational Biology. Public Library of Science; 2016;12: e1005140. doi:10.1371/journal.pcbi.1005140

113. Zhang C, Ji B, Mardinoglu A, Nielsen J, Hua Q. Logical transformation of genome-scale metabolic models for gene level applications and analysis. Bioinformatics. 2015;31: 2324–2331. doi:10.1093/bioinformatics/btv134

114. Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, et al. BRENDA in 2017: New perspectives and new tools in BRENDA. Nucleic Acids Research. 2017;45: D380–D388. doi:10.1093/nar/gkw952

115. Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, et al. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics (Oxford, England). Oxford University Press; 2017;33: 3454–3460. doi:10.1093/bioinformatics/btx439

116. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. Oxford University Press; 2017;45: D353–D361. doi:10.1093/nar/gkw1092

117. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, et al. The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry. 2011;50: 4402–4410. doi:10.1021/bi2002289

118. Tyson CB, Lord PG, Wheals AE. Dependency of size of Saccharomyces cerevisiae cells on growth rate. Journal of Bacteriology. 1979;138: 92–98.

119. Van Dijken JP, Bauer J, Brambilla L, Duboc P, Francois JM, Gancedo C, et al. An interlaboratory

comparison of physiological and genetic properties of four Saccharomyces cerevisiae strains. Enzyme and Microbial Technology. 2000;26: 706–714. doi:10.1016/S0141-0229(00)00162-9

120. De Deken RH. The Crabtree Effect: A Regulatory System in Yeast. Journal of General Microbiology. Microbiology Society; 1966;44: 149–156. doi:10.1099/00221287-44-2-149

121. Molenaar D, van Berlo R, de Ridder D, Teusink B. Shifts in growth strategies reflect tradeoffs in cellular economics. Molecular systems biology. Nature Publishing Group; 2009;5: 323. doi:10.1038/msb.2009.82

122. Niebel B, Leupold S, Heinemann M. An upper limit in Gibbs energy dissipation governs cellular metabolism. Nature Metabolism. Nature Publishing Group; 2018;in press: 125–132. doi:10.1038/s42255-018-0006-7

123. O'Brien EJ, Lerman J a, Chang RL, Hyduke DR, Palsson BØ. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Molecular systems biology. 2013;9: 693. doi:10.1038/msb.2013.52

124. Van Hoek P, Van Dijken JP, Pronk JT. Effect of specific growth rate on fermentative capacity of baker's yeast. Applied and Environmental Microbiology. 1998;64: 4226–4233.

125. Famili I, Forster J, Nielsen J, Palsson BØ. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proceedings of the National Academy of Sciences of the United States of America. National Acad Sciences; 2003;100: 13134–9. doi:10.1073/pnas.2235812100

126. Hukelmann JL, Anderson KE, Sinclair L V, Grzes KM, Murillo AB, Hawkins PT, et al. The cytotoxic T cell proteome and its shaping by the kinase mTOR. Nature Immunology. Nature Publishing Group; 2016;17: 104–112. doi:10.1038/ni.3314

127. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics. Nature Publishing Group; 2010. pp. 733–739. doi:10.1038/nrg2825

128. Arike L, Valgepea K, Peil L, Nahku R, Adamberg K, Vilu R. Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli. Journal of Proteomics. 2012;75: 5437–5448. doi:10.1016/j.jprot.2012.06.020

129. Wiśniewski JR, Rakus D. Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the Escherichia coli proteome. Journal of Proteomics. 2014;109: 322–331. doi:10.1016/j.jprot.2014.07.012

130. Wiśniewski JR, Ostasiewicz P, Duś K, Zielińska DF, Gnad F, Mann M. Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. Molecular Systems Biology. 2012;8. doi:10.1038/msb.2012.44

131. Karp NA, Huber W, Sadowski PG, Charles PD, Hester S V, Lilley KS. Addressing Accuracy and Precision Issues in iTRAQ Quantitation. Molecular & Cellular Proteomics. American Society for Biochemistry and Molecular Biology; 2010;9: 1885–1897. doi:10.1074/mcp.M900628-MCP200

132. Jenner L, Melnikov S, de Loubresse NG, Ben-Shem A, Iskakova M, Urzhumtsev A, et al. Crystal structure of the 80S yeast ribosome. Current Opinion in Structural Biology. Elsevier Current Trends; 2012. pp. 759–767. doi:10.1016/j.sbi.2012.07.013

133. Fabre B, Lambour T, Bouyssié D, Menneteau T, Monsarrat B, Burlet-Schiltz O, et al. Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry. EuPA Open Proteomics. Elsevier; 2014;4: 82–86. doi:10.1016/j.euprot.2014.06.001

134. Escobar-Henriques M, Daignan-Fornier B. Transcriptional regulation of the yeast GMP synthesis pathway by its end products. Journal of Biological Chemistry. 2001;276: 1523–1530. doi:10.1074/jbc.M007926200

135. Caspeta L, Chen Y, Ghiaci P, Feizi A, Baskov S, Hallström BM, et al. Altered sterol composition renders yeast thermotolerant. Science. 2014;346: 75–78. doi:10.1126/science.1258137

136. Loman N, Watson M. So you want to be a computational biologist? Nature Biotechnology. Nature Publishing Group; 2013;31: 996–998. doi:10.1038/nbt.2740

137. Bergenholm D, Gossing M, Wei Y, Siewers V, Nielsen J. Modulation of saturation and chain length of fatty acids in Saccharomyces cerevisiae for production of cocoa butter-like lipids. Biotechnology and Bioengineering. Wiley-Blackwell; 2018;115: 932–942. doi:10.1002/bit.26518

138. Tiukova IA, Prigent S, Nielsen J, Sandgren M, Kerkhoven EJ. Genome-scale model of Rhodotorula toruloides metabolism. bioRxiv. Cold Spring Harbor Laboratory; 2019; 528489. doi:10.1101/528489

139. Massaiu I, Pasotti L, Sonnenschein N, Rama E, Cavaletti M, Magni P, et al. Integration of enzymatic data in Bacillus subtilis genome-scale metabolic model improves phenotype predictions and enables in silico design of poly-γ-glutamic acid production strains. Microbial Cell Factories. BioMed Central; 2019;18: 3. doi:10.1186/s12934-018-1052-2

140. Nilsson A, Nielsen J, Palsson BO. Metabolic Models of Protein Allocation Call for the Kinetome. Cell Systems. Cell Press; 2017. pp. 538–541. doi:10.1016/j.cels.2017.11.013

141. Ho B, Baryshnikova A, Brown GW. Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. Cell Systems. Elsevier; 2018;6: 192–205.e3. doi:10.1016/j.cels.2017.12.004

142. Vowinckel J, Zelezniak A, Bruderer R, Mülleder M, Reiter L, Ralser M. Cost-effective generation of precise label-free quantitative proteomes in high-throughput by microLC and data-independent acquisition. Scientific Reports. 2018;8: 1–10. doi:10.1038/s41598-018-22610-4

143. Frainay C, Schymanski EL, Neumann S, Merlet B, Salek RM, Jourdan F, et al. Mind the gap: Mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. Metabolites. Multidisciplinary Digital Publishing Institute; 2018;8: 51. doi:10.3390/metabo8030051

144. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models. PLoS Computational Biology. Public Library of Science; 2014;10: e1003882. doi:10.1371/journal.pcbi.1003882

145. Box GEP. Science and statistics. Journal of the American Statistical Association. 1976;71: 791–799. doi:10.1080/01621459.1976.10480949

146. Babtie AC, Stumpf MPH. How to deal with parameters for whole-cell modelling. Journal of the Royal Society Interface. 2017;14. doi:10.1098/rsif.2017.0237

147. Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, et al. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. Proceedings of the National Academy of Sciences of the United States of America. 2016;113: 3401–6. doi:10.1073/pnas.1514240113

148. Goel A, Eckhardt TH, Puri P, de Jong A, Branco dos Santos F, Giera M, et al. Protein costs do not explain evolution of metabolic strategies and regulation of ribosomal content: Does protein investment explain an anaerobic bacterial Crabtree effect? Molecular Microbiology. 2015;97: 77–92. doi:10.1111/mmi.13012

149. Noor E, Flamholz A, Bar-Even A, Davidi D, Milo R, Liebermeister W. The Protein Cost of Metabolic Fluxes: Prediction from Enzymatic Rate Laws and Cost Minimization. PLoS Computational Biology. 2016;12: 1–47. doi:10.1371/journal.pcbi.1005167

150. Liu JK, O'Brien EJ, Lerman JA, Zengler K, Palsson BO, Feist AM. Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. BMC systems biology. 2014;8: 110. doi:10.1186/s12918-014-0110-6